Journal of Machine Learning Research 22 (2021) 1-41 Submitted 9/20; Revised 4/21; Published 5/21

Learning a High-dimensional Linear Structural Equation Model via ℓ_1-Regularized Regression

Gunwoong Park GW.PARK23@GMAIL.COM
Department of Statistics
University of Seoul
Seoul, 02504, South Korea

Sang Jun Moon KENU1@NAVER.COM
Department of Statistics
University of Seoul
Seoul, 02504, South Korea

Sion Park ROCKGOAT95@GMAIL.COM
Department of Statistics
University of Seoul
Seoul, 02504, South Korea

Jong-June Jeon JJ.JEON@GMAIL.COM
Department of Statistics
University of Seoul
Seoul, 02504, South Korea

Editor: Victor Chernozhukov

Abstract

This paper develops a new approach to learning high-dimensional linear structural equation models (SEMs) without the commonly assumed faithfulness, Gaussian error distribution, and equal error distribution conditions. A key component of the algorithm is component-wise ordering and parent estimations, where both problems can be efficiently addressed using ℓ_1-regularized regression. This paper proves that sample sizes $n = \Omega(d^2 \log p)$ and $n = \Omega(d^2 p^{2/m})$ are sufficient for the proposed algorithm to recover linear SEMs with sub-Gaussian and $(4m)$-th bounded-moment error distributions, respectively, where p is the number of nodes and d is the maximum degree of the moralized graph. Further shown is the worst-case computational complexity $O(n(p^3 + p^2 d^2))$, and hence, the proposed algorithm is statistically consistent and computationally feasible for learning a high-dimensional linear SEM when its moralized graph is sparse. Through simulations, we verify that the proposed algorithm is statistically consistent and computationally feasible, and it performs well compared to the state-of-the-art US, GDS, LISTEN and TD algorithms with our settings. We also demonstrate through real COVID-19 data that the proposed algorithm is well-suited to estimating a virus-spread map in China.

Keywords: bayesian networks, causal learning, directed acyclic graph, linear structural equation model, structure learning, ℓ_1-regularization

1. Introduction

Directed acyclic graphical (DAG) models, also referred to as Bayesian networks, are popular probabilistic statistical models that are associated with a graph where nodes correspond to variables of interest. In addition, the edges of the graph describe conditional dependence information and causal/directional relationships among the variables. Hence, the models have been applied to a large number of fields, including bioinformatics, social science, control theory, image processing, and marketing analysis, among others (see e.g., Hausman, 1983; Newey et al., 1999; Friedman et al., 2000; Imbens and Newey, 2009; Nagarajan et al., 2013; Peters and Bühlmann, 2014). However, structure learning for graphical models from the observational distribution remains an open challenge due to *non-identifiability* and double-exponentially growing *computational complexity* in the number of nodes. Therefore, learning DAG models from purely observational data, also known as a causal discovery, has drawn much attention.

Recently, it has been shown that DAG models can be fully identifiable with some restrictions on the distributions. For example, Shimizu et al. (2006); Zhang and Hyvärinen (2009b) show that linear non-Gaussian additive noise models are identifiable where each variable is determined by a linear function of its parents plus an independent error term. Zhang and Hyvärinen (2009a); Hoyer et al. (2009); Mooij et al. (2009); Peters et al. (2012) discuss the identifiability of non-linear SEMs where each variable is determined by a non-linear function of its parents and an error term. Park and Raskutti (2015, 2018) prove the identifiability of DAG models where a conditional distribution of each node given its parents belongs to some exponential family distributions. Lastly, Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Park and Kim (2020) prove that linear SEMs can be identifiable using error variances. We refer the readers to Peters et al. (2014); Eberhardt (2017); Glymour et al. (2019); Park (2020) for their detailed review.

For these identifiable DAG models, recent literature has developed statistically consistent learning algorithms that can be executed under the identifiability condition to target the identifiable class of DAG models. Furthermore, the algorithms focus on learning the models in polynomial time, as well as in a high-dimensional regime (e.g., Shimizu et al., 2011; Bühlmann et al., 2014; Ghoshal and Honorio, 2017; Park and Park, 2019a; Chen et al., 2019; Park and Kim, 2021).

In terms of learning high-dimensional (sub-)Gaussian linear SEMs, Ghoshal and Honorio (2018) develops a constrained ℓ_1-minimization for inverse covariance matrix estimation (CLIME)-based algorithm with the sample bound $n = \Omega(d^4 \log p)$, in which d is the maximum degree of the moralized graph when error variances are heterogeneous. Furthermore, in the equal error variance setting, Loh and Bühlmann (2014); Ghoshal and Honorio (2017) provide graphical lasso-based approaches with sample complexity $n = \Omega(d^4 \log p)$. In addition, Chen et al. (2019) develops a best subset-based algorithm with sample complexity $n = \Omega(q^2 \log p)$ for ordering estimation, in which q is the predetermined upper bound of the maximum indegree (see the brief review in Section 2.3). However, these existing algorithms have not yet focused on a regularized regression-based approach for learning a high-dimensional linear SEM, whereas undirected graphical models have been successfully estimated based on ℓ_1-regularized regression.

This paper focuses on developing a new ℓ_1-regularized regression-based algorithm for learning high-dimensional linear SEMs, which allows a broader class of error distribution having the $(4m)$-th bounded moment. Also proven are its sample complexities $n = \Omega(d^2 \log p)$ and $n = \Omega(d^2 p^{2/m})$, under which the proposed algorithm recovers linear SEMs with sub-Gaussian and $(4m)$-th bounded-moment error distributions, respectively. Further shown is that the proposed algorithm is computationally polynomial $O(np^3 + nd^2p^2)$ in the worst case. Lastly, it is pointed out that the algorithm does not require the commonly used faithfulness, Gaussian error distribution, and equal error distribution assumptions that might be very restrictive. Section 3.3 provides comparisons in both computational and sample complexities between the proposed and the high-dimensional linear SEM learning algorithms, which are the linear structural equation model learning (LISTEN) (Ghoshal and Honorio, 2018) and the top-down search (TD) (Chen et al., 2019) algorithms.

We demonstrate through simulations and real COVID-19 data that the proposed algorithm performs well in terms of recovering directed edges. In the simulation study, we consider low- and high-dimensional linear SEMs where the number of nodes is $p \in \{25, 50, 100, 150, 200, 250\}$ and the maximum degree is $d \in \{5, 8\}$. Furthermore, the proposed algorithm is compared to the state-of-the-art uncertainty scoring (US) (Park, 2020), greedy DAG search (GDS) (Peters and Bühlmann, 2014), LISTEN, and TD algorithms.

The remainder of this paper is structured as follows. Section 2.1 summarizes the necessary notations and problem settings. Section 2.2 explains basic concepts of a linear SEM and Section 2.3 discusses identifiability conditions and existing learning algorithms. In Section 3, we introduce the new algorithm for high-dimensional linear SEM learning with sub-Gaussian and bounded-moment error distributions. Sections 3.1 and 3.2 provide the theoretical guarantees and computational complexity,respectively, of the proposed algorithm. Furthermore, Section 3.3 compares the proposed algorithm to relevant methods in terms of sample and computational complexities. Sections 4 and 5 evaluate our method and state-of-the-art algorithms using synthetic and real COVID-19 data. Lastly, Section 6 offers a discussion, and suggests future works.

2. Preliminaries

First introduced are some necessary notations and definitions for DAG models. Then, we give a detailed description of identifiability conditions and existing algorithms for learning linear SEMs.

2.1 Problem Set-up and Notations

Directed acyclic graph $G = (V, E)$ consists of a set of nodes $V = \{1, 2, ..., p\}$ and a set of directed edges $E \subset V \times V$ with no directed cycles. A directed edge from node j to k is denoted by (j, k) or $j \to k$. The set of *parents* of node k, denoted by $\mathrm{Pa}(k)$, consists of all nodes j such that $(j, k) \in E$. In addition, the set of *children*, denoted by $\mathrm{Ch}(j)$, consists of all nodes k such that $(j, k) \in E$. The set of *neighbors* of node j, denoted by $\mathrm{Ne}(j)$, consists of all nodes k connected by an edge. If there is a directed path $j \to \cdots \to k$, then k is called a *descendant* of j, and j is called an *ancestor* of k. The sets $\mathrm{De}(k)$ and $\mathrm{An}(k)$ denote the set of all descendants and ancestors, respectively, of node k. An important property of DAGs is that there exists a (possibly non-unique) *ordering* $\pi = (\pi_1, \pi_2, ..., \pi_p)$ of a directed graph that represents directions of edges such that for every directed edge $(j, k) \in E$, j comes before k in the ordering. Hence, learning a graph is equivalent to learning the ordering as well as its parents. Similar definitions and notations can be found in Lauritzen (1996); Spirtes et al. (2000).

We consider a set of random variables $X := (X_j)_{j \in V}$ with a probability distribution taking values in a sample space $\mathcal{X}_V$ over the nodes in G. Suppose that a random vector X has a joint probability density function $\Pr(G) = \Pr(X_1, X_2, ..., X_p)$. For any subset S of V, let $X_S := \{X_j : j \in S \subset V\}$ and $\mathcal{X}_S := \times_{j \in S} \mathcal{X}_j$ where $\mathcal{X}_j$ is a sample space of X_j. For any node $j \in V$, $\Pr(X_j \mid X_S)$ denotes the conditional distribution of a variable X_j given a random vector X_S. Then, a DAG model has the following factorization (Lauritzen, 1996):

$$\Pr(G) = \Pr(X_1, X_2, ..., X_p) = \prod_{j=1}^{p} \Pr(X_j \mid X_{\mathrm{Pa}(j)}), \tag{1}$$

where $\Pr(X_j \mid X_{\mathrm{Pa}(j)})$ is the conditional distribution of X_j given its parents variables $X_{\mathrm{Pa}(j)} := \{X_k : k \in \mathrm{Pa}(j) \subset V\}$.

This paper assumes that there are n independent and identically distributed (i.i.d.) samples $X^{1:n} := (X^{(i)})_{i=1}^{n}$ from a given graphical model where $X_{1:p}^{(i)} := (X_j^{(i)})_{j=1}^{p}$ is a p-variate random vector. The notation $\widehat{\ }$ denotes an estimate based on samples $X^{1:n}$. This paper also accepts the sparse moralized graph assumption and the causal sufficiency assumption of $X_{\mathrm{Pa}(j)}$ being the only source of confounding for X_j.

Lastly, an important concept this paper needs to introduce is the *moral* graph or the undirected graphical model representation of a DAG. Moralized graph G' for DAG $G = (V, E)$ is an undirected graph where $G' = (V, E')$ in which E' includes the edge set E for DAG G with directions removed plus edges between any nodes that are parents of a common child. Hence, the maximum degree of the moralized graph is always greater than or equal to the maximum indegree.

2.2 Linear Structural Equation Models

A linear SEM is a special DAG model where all errors are additive, and each variable is modeled as a linear function of its parents. Hence, it can be written in the following matrix form:

$$(X_1, X_2, ..., X_p)^\top = B(X_1, X_2, ..., X_p)^\top + (\epsilon_1, \epsilon_2, ..., \epsilon_p)^\top, \tag{2}$$

where $B \in \mathbb{R}^{p \times p}$ is an edge weight matrix, an auto regression matrix, or a weighted adjacency matrix, with each element $[B]_{jk} = \beta_{jk}$, in which β_{jk} is the linear weight of an edge from X_k to X_j. Without loss of generality, this paper assumes that $\mathbb{E}(X_j) = 0$ for all $j \in V$. Then, the covariance matrix of the linear SEM and its inverse, referred to as Σ and Ω, respectively, are as follows:

$$\Sigma = (I_p - B)^{-1}\Sigma_\epsilon(I_p - B)^{-\top}, \text{ and } \quad \Theta = (I_p - B)^\top \Sigma_\epsilon^{-1}(I_p - B), \tag{3}$$

where $I_p \in \mathbb{R}^{p \times p}$ is the identity matrix, and $\Sigma_\epsilon = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_p^2)$ is a covariance matrix of the independent errors.

As a special case, where all errors are Gaussian in Equation (2), the joint density function of the model is as follows:

$$f_G(x_1, x_2, ..., x_p; \Theta) = \frac{1}{\sqrt{(2\pi)^p \det(\Theta^{-1})}} \exp\Big(-\frac{1}{2}(x_1, x_2, ..., x_p)\Theta(x_1, x_2, ..., x_p)^\top \Big). \quad (4)$$

Since the density function is parameterized by the inverse covariance matrix, recent Gaussian linear SEM learning approaches exploit various inverse covariance matrix estimation methods such as graphical lasso and CLIME (Loh and Bühlmann, 2014; Ghoshal and Honorio, 2017, 2018).

A natural extension of Gaussian linear SEMs is a sub-Gaussian linear SEM in which each error variable is sub-Gaussian; that is, $\epsilon_j/\sqrt{\mathrm{Var}(\epsilon_j)}$ is sub-Gaussian with parameter s_j. In a similar manner, a bounded-moment linear SEM is defined as the linear SEM with errors having a bounded-moment where $\max_{j \in V} \mathbb{E}((\epsilon_j/\sqrt{\mathrm{Var}(\epsilon_j)})^{4m}) \leq K_m$, where $K_m > 0$ only depends on m.

In the linear SEM setting, each variable X_j can be expressed as the following linear combination of independent errors corresponding to its ancestors:

$$X_j = \sum_{k \in \mathrm{Pa}(j)} \beta_{jk} X_k + \epsilon_j = \sum_{k \in \mathrm{An}(j)} \beta_{k \to j} \epsilon_k + \epsilon_j,$$

where $\beta_{k \to j}$ is the sum over products of coefficients along directed paths from k to j. Hence, if error variables hold a sub-Gaussian or a bounded-moment property, then X_j also satisfies a sub-Gaussian or a bounded-moment property.

2.3 Identifiability and Existing Algorithms

This section reviews recent works on learning linear SEMs. As discussed, Peters and Bühlmann (2014); Ghoshal and Honorio (2018); Park and Kim (2020) provide the following different identifiability conditions for linear SEMs, expressed in Equation (2), using error variances.

Lemma 1 (Identifiability Conditions in Theorem 2 of Park, 2020) *Consider a linear SEM (2) with DAG G and true ordering π. Then, DAG G is uniquely identifiable if either of the following conditions is satisfied.*

(a) Equal error variance condition:

$$\sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2.$$

(b) Forward stepwise selection condition; for any node $j = \pi_r \in V$ and $k \in De(j)$:

$$\begin{aligned} \sigma_j^2 &= Var(X_j \mid X_{\pi_1}, ..., X_{\pi_{r-1}}) \\ &< Var(X_k \mid X_{\pi_1}, , ..., X_{\pi_{r-1}}) = \sigma_k^2 + \mathbb{E}(Var(\mathbb{E}(X_k \mid X_{Pa(k)}) \mid X_{\pi_1}, ..., X_{\pi_{r-1}})). \end{aligned}$$

(c) Backward stepwise selection condition; for any node $j = \pi_r \in V$ and $\ell \in An(j)$:

$$\begin{aligned} \sigma_j^2 &= Var(X_j \mid X_{\pi_1}, ..., X_{\pi_r} \setminus X_j) \\ &> Var(X_\ell \mid X_{\pi_1}, ..., X_{\pi_r} \setminus X_\ell) = \sigma_\ell^2 - \mathbb{E}(Var(\mathbb{E}(X_\ell \mid X_{\pi_1}, ..., X_{\pi_r} \setminus X_\ell) \mid X_{Pa(\ell)})). \end{aligned}$$

In many areas, these conditions are acceptable and widely used; for example, the assumption of the exact same error variances, proposed in Peters and Bühlmann (2014), is used for applications with variables from a similar domain, a spatial or a time-series data. Hence, recent works developed linear SEM learning algorithms in three different settings: (i) Gaussian and non-Gaussian error distributions, (ii) low- and high-dimensional regimes, and (iii) homogeneous and heterogeneous error variances.

The GDS algorithm proposed in Peters and Bühlmann (2014) is a popular graph-wise estimation approach for a Gaussian linear SEM with homogeneous error variances. It applies the following penalized maximum likelihood by assuming all error distributions are Gaussian with the same error variances:

$$\{\widehat{B}, \widehat{\sigma}^2\} = \underset{B \in \mathbb{R}^{p \times p}, \sigma^2 \in \mathbb{R}^+}{\arg\min} \frac{np}{2} \log\left(2\pi\sigma^2\right) + \frac{n}{2\sigma^2} \mathrm{tr}\left\{(I-B)^\top (I-B)\widehat{\Sigma}\right\} + \frac{\log(n)}{2}\|B\|_0,$$

where n is the sample size, B is the weight matrix, and $\widehat{\Sigma}$ is the sample covariance matrix for X. In addition, $\|B\|_0 = |\{(j,k) \mid [B]_{jk} \neq 0\}|$. In principle, the GDS algorithm directly finds a directed graph, and hence, it calculates the likelihood of all possible graphs, where the number of graphs exponentially grows with the number of nodes. Hence, the drawback with the graph-wise estimation approach is an exponentially growing computational cost.

The TD algorithm developed in Chen et al. (2019) is a two-stage algorithm for learning high-dimensional sub-Gaussian linear SEMs with the same error variances. The first stage estimates an element of the ordering from the beginning by comparing the conditional variance. And then, the second stage performs parent estimations using existing variable selection techniques (Shojaie and Michailidis, 2010). More precisely, the r-th element of the ordering is determined using a best subset-based method with predetermined size q such that

$$\widehat{\pi}_r = \underset{j \in V \setminus \widehat{\pi}_{1:(r-1)}, S \subset \widehat{\pi}_{1:(r-1)}, |S|=q}{\arg\min} \mathrm{Var}(X_j \mid X_S).$$

Like the GDS algorithm, the TD algorithm may not be computationally feasible for a large-scale graph estimation, because it requires $O(p^{q+1})$ conditional variance estimations to infer the ordering. Furthermore, the specified q should be the upper bound of the maximum indegree; otherwise, there is no guarantee that the algorithm finds the true graph. However, this constraint enables the TD algorithm to recover a graph in a high-dimensional setting.

The GDS and TD algorithms assume the same error variances, and hence, the Ghoshal and Honorio (2018); Park and Kim (2020); Park (2020) develop structure learning algorithms based on the forward and backward stepwise selection conditions in Lemma 1. The US algorithm in Park (2020) focuses on learning a low-dimensional SEM with heterogeneous error variances. In other words, the algorithm can learn not only Gaussian linear SEMs, but also non-Gaussian and non-linear SEMs. More precisely, the algorithm first estimates an element of the ordering either from the beginning or the end using node-wise uncertainty scores. And then, it finds each directed edge using a conditional independence test. However, it is only applicable to the low-dimensional model.

The LISTEN algorithm proposed in Ghoshal and Honorio (2018) focuses on learning a high-dimensional linear SEM with heterogeneous error variances using the backward stepwise selection condition. It allows a broader class of error distributions with a sub-Gaussian and $(4m)$-th bounded moment. More specifically, the LISTEN algorithm first estimates the last element of the ordering using the diagonal entries of the inverse covariance matrix. And then, it determines its parents with non-zero entries on its row of the inverse covariance matrix. After eliminating the last element of the ordering, the algorithm applies the same procedure until a graph is completely estimated.

The existing algorithms have not yet focused on the ℓ_1-regularized regression-based approach for learning a high-dimensional linear SEM, whereas undirected graphical models have been successfully estimated based on ℓ_1-regularized regression. In particular, Meinshausen and Bühlmann (2006); Yang et al. (2015) establish consistency in learning (Gaussian) undirected graphical models via ℓ_1-regularized regression where the sample bound is $\Omega(d \log p)$ in which d is the degree of the undirected graph. This motivates a new regression-based algorithm that learns high-dimensional SEMs with relaxed error distributions holding the sub-Gaussian and bounded-moment properties. We provide details on the new algorithm in the next section.

3. Algorithm

This section introduces a new regression-based algorithm for high-dimensional linear SEMs. The proposed algorithm mainly exploits the backward stepwise selection condition, and hence, it allows heterogeneous error variances. A key component of the algorithm is component-wise ordering and parent estimations, where the problems can be efficiently addressed using ℓ_1-regularized regression. The overall process of the proposed algorithm is summarized in Algorithm 1.

To be specific, the proposed algorithm first estimates the last element of the ordering, and then determines its parents. The algorithm then estimates the next element of the ordering and its parents. It iterates this procedure until the complete graph structure is determined. Hence, the r-th iteration of the algorithm estimates π_{p+1-r} and its parents, given the estimated $\widehat{\pi}_{p+2-r}, ..., \widehat{\pi}_p$ if $r \geq 2$. More precisely, the r-th iteration of the algorithm is first conducted by following ℓ_1-regularized regression: For each node $j \in V$,

$$\widehat{\theta}_j(r) := \underset{\theta \in \mathbb{R}^{|S_j(r)|}}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} (X_j^{(i)} - \langle X_{S_j(r)}^{(i)}, \theta \rangle)^2 + \lambda \|\theta\|_1, \tag{5}$$

where $S_j(r) = V \setminus (\{j\} \cup \{\widehat{\pi}_{p+2-r}, ..., \widehat{\pi}_p\})$ if $r \geq 2$; otherwise, $S_j(r) = V \setminus \{j\}$. In addition, $\langle \cdot, \cdot \rangle$ represents the inner product. This can be understood as the neighborhood estimation of node j in the moralized graph under the faithfulness assumption where each conditional independence relationship between X_j and X_k given X_S is equivalent to d-separation between j and k given S (see details in Spirtes et al., 2000). However, it should be pointed out that, without the faithfulness assumption, it still uncovers the parents for π_{p+1-r} and the subset of the neighborhood for $j \in \{\pi_1, ..., \pi_{p-r}\}$ in the population. The detailed justification is provided in Section 3.1.

Then, the algorithm determines π_{p+1-r} with the largest conditional variance, given $X_{S_j(r)}$ for all $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, using the backward stepwise selection condition in Lemma 1. Based on the result of ℓ_1-regularized regression expressed in Equation (5), there are various consistent estimators, such as the residual sum of squares estimator, the cross-validation-based estimator, and the refitted cross-validation estimator (see more details in

Fan et al., 2012). In principle, any consistent estimator can be applied; however, this paper focuses on the following two-stage ℓ_1-regularization method:

$$\widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)}) := \frac{1}{n}\sum_{i=1}^{n}\left(X_j^{(i)} - \langle X_{\widehat{T}_j(r)}^{(i)}, \widehat{\alpha}_j(r)\rangle\right)^2, \tag{6}$$
$$\text{where} \quad \widehat{\alpha}_j(r) := \arg\min_{\alpha \in \mathbb{R}^{|\widehat{T}_j(r)|}} \frac{1}{n}\sum_{i=1}^{n}\left(X_j^{(i)} - \langle X_{\widehat{T}_j(r)}^{(i)}, \alpha\rangle\right)^2,$$

and $\widehat{T}_j(r)$ is the support of $\widehat{\theta}_j(r)$ (i.e., $\widehat{T}_j(r) := \{k \in S_j(r) \mid [\widehat{\theta}_j(r)]_k \neq 0\}$) in which $[\widehat{\theta}_j(r)]_k$ is an element of $\widehat{\theta}_j(r)$ corresponding to the random variable X_k. In addition, $\widehat{\alpha}_j(r)$ is the ordinary least square estimator where X_j is a response variable and $X_{\widehat{T}_j(r)}$ are explanatory variables.

The parent estimation for π_{p+1-r} is direct from the solution of ℓ_1-regularized regression, because its support is the parents for π_{p+1-r} in the population (see Proposition 2). Hence, the parents of node $j = \pi_{p+1-r}$ are determined as $\widehat{\mathrm{Pa}}(j) := \{k \in S_j(r) : [\widehat{\theta}_j(r)]_k \neq 0\}$ where $\widehat{\theta}_j(r)$ is the solution to Equation (5).

The proposed approach is essentially an optimization problem that guarantees a global optimum for the objective function in Equation (5). In other words, the performance of the proposed method highly depends on that of ℓ_1-regularized regression. Hence, it intuitively makes sense that the proposed algorithm consistently learns a high-dimensional linear SEM without the commonly used faithfulness, known degree, and Gaussian error distribution assumptions that are not necessary for ℓ_1-regularized regression. Furthermore, the proposed method is computationally as efficient as ℓ_1-regularized regression.

Algorithm 1: High-dimensional Linear SEM Learning Algorithm

```
Input  : $n$ i.i.d. samples, $X^{1:n}$
Output: Estimated graph structure, $\widehat{G} = (V, \widehat{E})$
Set $\widehat{\pi}_{p+1} = \emptyset$ ;
for $r = \{1, 2, ..., p-1\}$ do
    for $j \in V \setminus \{\widehat{\pi}_{p+1}, ..., \widehat{\pi}_{p+2-r}\}$ do
        $S_j(r) = V \setminus (\{j\} \cup \{\widehat{\pi}_{p+1}, ..., \widehat{\pi}_{p+2-r}\})$ ;
        Estimate $\widehat{\theta}_j(r)$ for $\ell_1$-regularized regression in Equation (5);
        Estimate conditional variances $\widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)})$ using Equation (6);
    end
    Determine the $(p+1-r)$-th element of the ordering:
     $\widehat{\pi}_{p+1-r} = \arg\max_j \widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)})$;
    Determine the parents of $\widehat{\pi}_{p+1-r}$: $\widehat{\mathrm{Pa}}(\widehat{\pi}_{p+1-r}) = \{k \in S_j(r) : [\widehat{\theta}_{\widehat{\pi}_{p+1-r}}(r)]_k \neq 0\}$;
end
Return: Estimate an edge set, $\widehat{E} = \cup_{r \in \{1,2,...,p-1\}}\{(k, \widehat{\pi}_{p+1-r}) : k \in \widehat{\mathrm{Pa}}(\widehat{\pi}_{p+1-r})\}$
```

3.1 Theoretical Guarantees

This section provides the statistical guarantees on Algorithm 1 for learning high-dimensional linear SEMs (2). As discussed, although the proposed algorithm runs with any appropriate estimator for conditional variances, we focus on the case where the two-stage lasso variance estimator (6) is applied.

We begin in Section 3.1.1 by stating the assumptions on the sample covariance matrix required in our analysis, including a particular type of mutual incoherence or irrepresentability condition. Then, in Section 3.1.2, we state our main results on the consistency of Algorithm 1 for both sub-Gaussian and $(4m)$-th bounded-moment linear SEMs. The main results are expressed in terms of the triple (n, p, d) where n is the sample size, p is the number of nodes, and d is the maximum degree of the moralized graph.

For ease of notation, $\theta_j^*(r)$ denotes the solution to Equation (5) when $\lambda = 0$ in the population. Then, it can be expressed as follows:

$$\theta_j^*(r) := \arg\min_{\theta \in \mathbb{R}^{|S_j^*(r)|}} \mathbb{E}\left((X_j - \langle X_{S_j^*(r)}, \theta \rangle)^2 \right), \tag{7}$$

where $S_j^*(r) = \{\pi_1, ..., \pi_{p+1-r}\} \setminus \{j\}$ and $\langle \cdot, \cdot \rangle$ represents the inner product.

Simple algebra yields $\Sigma_{S_j^*(r)S_j^*(r)} \theta_j^*(r) = \Sigma_{S_j^*(r)j}$ where $\Sigma_{S_j^*(r)S_j^*(r)}$ is a sub-matrix of the true covariance matrix Σ corresponding to variables $X_{S_j^*(r)}$. Hence, we re-formalize X_j as follows:

$$X_j = \langle X_{\mathrm{Pa}(j)}, \beta_j^* \rangle + \epsilon_j = \langle X_{S_j^*(r)}, \theta_j^*(r) \rangle + \delta_j,$$

where $\beta_j^* = (\beta_{jk})_{k \in \mathrm{Pa}(j)}$ in Equation (2), and $\delta_j = X_j - \langle X_{S_j^*(r)}, \theta_j^*(r) \rangle$.

In the special case where j is π_{p+1-r}, $\theta_j^*(r)$ corresponds exactly to the set of true parameters; that is, $[\theta_j^*(r)]_k = \beta_{jk}$ if $k \in \mathrm{Pa}(j)$; otherwise, $[\theta_j^*(r)]_k = 0$. However, our results apply more generally for $j \neq \pi_{p+1-r}$. Hence, the following proposition is necessary to guarantee that the above re-formalized problem in terms of $\theta_j^*(r)$ is a still sparse regression problem under the bounded degree of the moralized graph condition.

Proposition 2 *For any $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, the true support of the solution $\theta_j^*(r)$ to Equation (7) is a subset of the neighborhood of j in the moralized graph:*

$$Supp(\theta_j^*(r)) \subset Ne(j).$$

In addition, for $j = \pi_{p+1-r}$, it is the parents of j:

$$Supp(\theta_j^*(r)) = Pa(j).$$

The detailed proof is provided in Appendix C.1. It should be pointed out that Proposition 2 does not require the faithfulness assumption that can be very restrictive especially for linear SEMs (see details in Uhler et al., 2013). The faithfulness assumption guarantees that $\mathrm{Supp}(\theta_j^*(r)) = \mathrm{Ne}(j)$ for $j \in \{\pi_1, ..., \pi_{p+1-r}\}$. However, our focus is to learn the parents of $j = \pi_{p+1-r}$, and hence, Algorithm 1 accurately recovers a graph without the faithfulness assumption.

3.1.1 Assumptions

We begin by discussing the assumptions we impose on the linear SEM (2) for both sub-Gaussian and bounded-moment error distributions. Our first two assumptions are prevalent in the literature, such as Wainwright et al. (2006); Ravikumar et al. (2011); Yang et al. (2015); Park and Raskutti (2018); Park and Park (2019b) where ℓ_1-regularized regression was used for graphical model learning.

Assumption 3 (Incoherence Assumption) *For any $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, there exists a positive constant $\gamma > 0$ such that*

$$\max_{j,r} \max_{t \in T_j(r)^c} \left\| \sum_{i=1}^{n} (X_t^{(i)})^\top X_{T_j(r)}^{(i)} \left(\sum_{i=1}^{n} (X_{T_j(r)}^{(i)})^\top X_{T_j(r)}^{(i)} \right)^{-1} \right\|_\infty \leq 1 - \gamma,$$

where $T_j(r) \subset Ne(j)$ is the support of the solution $\theta_j^*(r)$ to Equation (7).

Assumption 4 (Dependency Assumption) *There exists a positive constant $\lambda_{\min} > 0$ such that*

$$\min_{j \in V} \Lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} (X_{Ne(j)}^{(i)})^\top X_{Ne(j)}^{(i)} \right) \geq \lambda_{\min},$$

where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A.

The incoherence assumption ensures that parent and non-parent variables are not overly correlated. Furthermore, the dependency condition forces the number of neighbors to be small ($d < n$), and ensures that variables belonging to the neighborhood do not become overly dependent.

Assumption 5 *For any node $j = \pi_r \in V$ and $\ell \in An(j)$, there exists a positive constant $\tau_{\min} > 0$ such that*

$$\sigma_j^2 - \sigma_\ell^2 + \mathbb{E}(Var(\mathbb{E}(X_\ell \mid X_{\pi_1}, ..., X_{\pi_r} \setminus X_\ell) \mid X_{Pa(\ell)})) > \tau_{\min}.$$

Assumption 6 (Minimum Signal Assumption) *For any $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, there exists a positive constant $\theta_{\min} > 0$ such that*

$$\min_{j,r} |\theta_j^*(r)| > \theta_{\min}.$$

Assumption 5 is the sample version of the backward selection identifiability condition in Lemma 1. Since this section focuses on learning a linear SEM in the finite sample setting, this stronger identifiability assumption is inevitable. In the same manner, Assumption 6 is required to ensure that each non-zero coefficient of $\theta_j^*(r)$ is sufficiently far away from zero.

Although our assumptions are standard in the previous graphical model learning approaches using ℓ_1-regularized regressions (e.g., Meinshausen and Bühlmann, 2006; Ravikumar et al., 2011; Yang et al., 2015), it should be noted that these assumptions might be very restrictive and non-checkable. For instance, the minimum signal assumption, also referred to as the beta-min condition, is questionable to be true in a real-world problem when the presence of weak signals cannot be ruled out.

However, we also point out that a lot of recent works have studied variable selection approaches under weaker assumptions (e.g., Bühlmann et al., 2013; Zhang and Zhang, 2014; Chernozhukov et al., 2019). In principle, accurate variable selection and conditional variance estimation are sufficient for recovering linear SEM, and thus, we believe that one can develop a more practical and consistent algorithm under milder assumptions. We leave this to future study.

3.1.2 Main Result

This section provides the theoretical results of Algorithm 1 in terms of the triple (n, p, d) where n is the sample size, p is the number of nodes, and d is the maximum degree of the moralized graph. More precisely, it provides theoretical guarantees on the ℓ_1-regularized regression problem in Equation (5) and on the ordering estimation problem when the two-stage lasso variance estimator (6) is applied.

Theorem 7 *Consider a linear SEM* (2) *with sub-Gaussian and* ($4m$)*-th bounded-moment errors. Suppose that Assumptions 3, 4, and 6 are satisfied. In addition, suppose that a regularization parameter is* $\lambda \in (0, \min\{\theta_{\min}\lambda_{\min}, \frac{10\sigma_{\max}^2}{\gamma}\})$ *where* $\sigma_{\max}^2$ *is the maximum error variance.*

- *For a sub-Gaussian linear SEM, there exist positive constants* C_1, C_2, C_3*, and* $C_4 > 0$ *such that, for any* $r \in \{1, 2, ..., p-1\}$ *and* $j \in \{\pi_1, ..., \pi_{p+1-r}\}$,

$$\Pr\left(sign\left(\widehat{\theta}_j(r)\right) = sign\left(\theta_j^*(r)\right)\right) \geq 1 - C_1 exp\left(\frac{-C_2 n\lambda^2}{d^2}\right) - C_3 exp\left(\frac{-C_4 n}{d^2}\left(\theta_{\min} - \frac{\lambda}{\lambda_{\min}}\right)^2\right).$$

- *For a* ($4m$)*-th bounded-moment linear SEM, there exist positive constants* D_1 *and* $D_2 > 0$ *such that, for any* $r \in \{1, 2, ..., p-1\}$ *and* $j \in \{\pi_1, ..., \pi_{p+1-r}\}$,

$$\Pr\left(sign\left(\widehat{\theta}_j(r)\right) = sign\left(\theta_j^*(r)\right)\right) \geq 1 - D_1\frac{d^{2m}}{n^m\lambda^{2m}} - D_2\frac{d^{2m}}{n^m}\left(\theta_{\min} - \frac{\lambda}{\lambda_{\min}}\right)^{-2m}.$$

The detailed proof is in Appendix A. The key technique for the proof is the *primal-dual witness* method used in sparse ℓ_1-regularized regressions and related techniques as in Meinshausen and Bühlmann (2006); Wainwright et al. (2006); Ravikumar et al. (2010); Yang et al. (2015). Theorem 7 intuitively makes sense, because neighborhood selection via ℓ_1-regularized regression is a well-studied problem, and its bias can be controlled by choosing an appropriate regularization parameter. Hence, our ℓ_1-regularized regression-based approach successfully recovers the sign of the solution in Equation (7) for sub-Gaussian and ($4m$)-th bounded-moment error variables.

A combination of Theorem 7 and Proposition 2 implies that the directed edges can be recovered with high probability when the ordering is provided. Hence, with an appropriate regularization parameter, and applying the union bound, if $n = \Omega(d^2 \log p)$ for a sub-Gaussian linear SEM and if $n = \Omega(d^2 p^{1/m})$ for a ($4m$)-th bounded-moment linear SEM, the proposed algorithm accurately learns the parents with high probability.

Theorem 8 *Consider a linear SEM* (2) *with sub-Gaussian and* (4*m*)*-th bounded-moment errors. Suppose that Assumptions 3, 4, 5, and 6 are satisfied. In addition, for any* $r \in \{1, 2, ..., p-1\}$ *and* $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, *the supports of* $(\theta_j^*(r))$ *are correctly recovered. Then, Algorithm 1 estimates the ordering with high probability.*

- *For a sub-Gaussian linear SEM, there exist positive constants* C_1 *and* $C_2 > 0$ *such that*

$$\Pr\left(\widehat{\pi} = \pi\right) > 1 - C_1 p^2 exp\left(\frac{-C_2 n}{d^2}\right).$$

- *For a* (4*m*)*-th bounded-moment linear SEM, there exists a positive constant* $D_1 > 0$ *such that*

$$\Pr\left(\widehat{\pi} = \pi\right) > 1 - D_1 p^2 \frac{d^{2m}}{n^m}.$$

The main point of the proof is to show the consistency of the two-stage lasso conditional variance estimator and the conditional variance comparisons based on the backward selection condition in Lemma 1. This also intuitively makes sense because, in principle, it involves the conditional variance estimations in low-dimensional settings where the number of variables belonging to the conditioning set is up to d. However, there are $p(p+1)/2 - 1$ conditional variances to be estimated. Furthermore, sub-Gaussian and $(4m)$-th bounded-moment variables are considered, and hence, we provide the detailed proof in Appendix B.

Finally, by combining Theorems 7, 8, and Proposition 2, we reach the final main result that Algorithm 1, with high probability, successfully recovers the true structure of a sub-Gaussian and a $(4m)$-th bounded-moment linear SEM, respectively.

Corollary 9 (Consistency of the Proposed Algorithm) *Consider a linear SEM* (2) *with sub-Gaussian and bounded-moment errors. Suppose that Assumptions 3, 4, 5, and 6 are satisfied, and an appropriate regularization parameter is chosen.*

- *For a sub-Gaussian linear SEM, Algorithm 1 finds the true graph with high probability if sample size* $n = \Omega(d^2 \log p)$.

- *For a* (4*m*)*-th bounded-moment linear SEM, Algorithm 1 finds the true graph with high probability if sample size* $n = \Omega(d^2 p^{2/m})$.

3.2 Computational Complexity

In this section, we provide the computational complexity for Algorithm 1 where the lasso variance estimator seen in Equation (6) is applied, as discussed in Section 3.1. Then, the proposed algorithm involves $O(p^2)$ ℓ_1-regularized regressions where the worst-case complexity is $O(np)$ for a single ℓ_1-regularized regression run (Friedman et al., 2009). Specifically, the coordinate descent method updates each gradient in $O(p)$ operations. Hence, with the maximum d non-zero terms in the regression in Equation (5), a complete cycle costs $O(pd)$ operations if no new variables become non-zero, and costs $O(np)$ for each new variable entered (see details in Friedman et al., 2010). Since Algorithm 1 has $p - 1$ iterations, and there are $p+1-r$ regressions with $p-r$ independent variables for the r-th iteration, the total worst-case complexity for ℓ_1-regularized regressions is $O(np^3)$.

For a conditional variance estimation, there are $p(p+1)/2-1$ conditional variance estimations, and the worst-case computational cost of each estimation is $O(nd^2+d^3)$. Since a sparse moralized graph is assumed and a comparison of conditional variances takes up to $O(p^2 \log p)$ for all $p-1$ iterations, the total worst-case complexity in recovering the ordering is $O(nd^2p^2+p^2\log p)$. Consequently, Algorithm 1 has a polynomial computational complexity of $O(np^3+nd^2p^2)$ at worst.

3.3 Comparisons to Other Works

Here, we compare our theoretical results against those in some related works. We first compare Algorithm 1, against the high-dimensional linear SEM learning LISTEN, and TD algorithms in terms of the sample complexity. Ghoshal and Honorio (2018) shows that the CLIME-based LISTEN algorithm successfully learns a sub-Gaussian linear SEM with high probability if the sample size is sufficiently large $n=\Omega(d^4\log p)$. In addition, Chen et al. (2019) proves that the best subset-based TD algorithm successfully learns a sub-Gaussian linear SEM with high probability if the sample size is sufficiently large $n=\Omega(q^2\log p)$, where q is the predetermined upper bound of the maximum indegree. Hence, it can be $n=\Omega(d_{in}^2\log p)$ where $d_{in}\le d$ is the maximum indegree of a graph. However, in terms of the sample complexity of ℓ_1-regularized-based Algorithm 1, Section 3.1.2 shows that it accurately estimates a sub-Gaussian linear SEM with high probability if the sample size scales at $n=\Omega(d^2\log p)$. Hence, the TD algorithm has the smaller sample complexity than the proposed and LISTEN algorithms when learning a sub-Gaussian linear SEM.

In terms of $(4m)$-th bounded-moment linear SEM learning, Ghoshal and Honorio (2018) shows that the LISTEN algorithm accurately recovers a graph if sample size $n=\Omega(d^4p^{2/m})$. In contrast, the proposed algorithm requires sample size $n=\Omega(d^2p^{2/m})$. Hence, in learning both sub-Gaussian and bounded-moment linear SEMs, the proposed algorithm has a better sample complexity than the LISTEN algorithm. This difference in sample complexity can also be found in Gaussian undirected graphical model learning between lasso-based learning and graphical lasso-based learning approaches.

Now, we compare the computational complexities of the proposed, LISTEN, and TD algorithms. The LISTEN algorithm has a complexity of $O(n(p^3+pd^4))$ in learning a sparse linear SEM. In addition, the TD algorithm requires $O(np^q)$ in which q is the upper bound of the maximum indegree of a graph. Lastly, the worst-case complexity of the proposed algorithm is $O(n(p^3+p^2d^2))$. We note that the computational complexities of the proposed and LISTEN algorithms rely on the maximum degree of the moralized graph, whereas that of the TD algorithm depends on the maximum indegree. Hence, they cannot be directly compared. However, it is expected that if a graph has a node with a lot of parent nodes such as a hub-node, the TD algorithm would require a huge run time. Comparisons of average run times for the algorithms are provided in Section 4.3.

Lastly, we compare our algorithm to the ℓ_1-regularized-based Gaussian undirected graphical model learning method. Meinshausen and Bühlmann (2006) shows that Gaussian undirected graphical models can be recovered via ℓ_1-regularized regression if sample size $n=\Omega\left(d\log p\right)$, where d is the degree of the undirected graph. In contrast, this paper shows that Gaussian linear SEMs can be learned via ℓ_1-regularized regression if $n = \Omega(d^2 \log p)$ where d is obtained by the moralized graph. Furthermore, in terms of computational complexity, the proposed algorithm is p times slower in the worst case. These differences in the sample and time complexities mainly come from the presence of the ordering.

4. Numerical Experiments

This section presents the empirical performance of Algorithm 1 to support our theoretical results indicating that the proposed algorithm consistently learns not only Gaussian linear SEMs, but general linear SEMs with light and heavy tail distributions under appropriate conditions. Hence, considered are high-dimensional (i) Gaussian linear SEMs and (ii) general linear SEMs where error distributions are sequentially Gaussian, Uniform, Student's t, and Beta. Also shown is a comparison of Algorithm 1, the US (Park and Kim, 2020), GDS (Peters and Bühlmann, 2014), LISTEN (Ghoshal and Honorio, 2018), and TD (Chen et al., 2019) algorithms in terms of both accuracy and computational cost.

The proposed algorithm was evaluated for the empirical probability of successfully recovering all edges; that is, $P(E = \widehat{E})$. In addition, the proposed algorithm and the comparison US, GDS, LISTEN, and TD algorithms were evaluated in terms of the average Hamming distance between the estimated and true DAGs (the number of edges that are different between two graphs). For the Hamming distance, smaller is better.

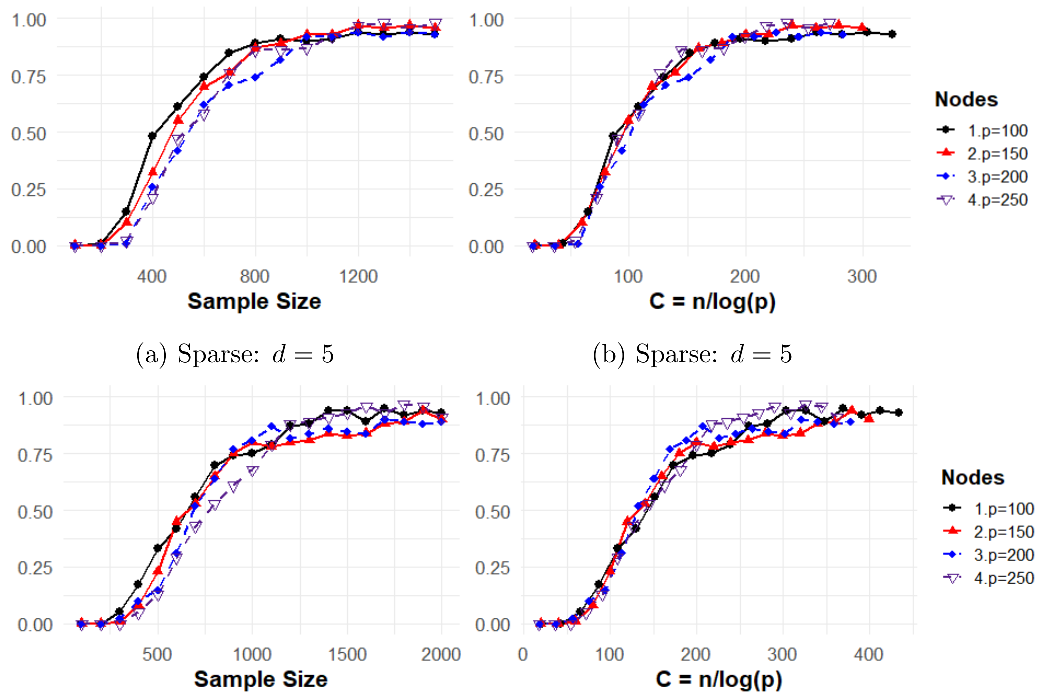
To validate our theoretical findings from Theorems 7 and 8, the two-stage lasso-based conditional variance estimator was applied to the proposed method. In terms of the regularization parameters for the proposed and LISTEN algorithms, they were set to $2\sqrt{\frac{\log p}{n}}$. In addition for the LISTEN algorithm, the hard threshold parameter was set to half of the minimum value of true edge weights, $\min(|\beta^*_{jk}|/2)$, by using the true model information. Lastly, for the TD algorithm, we always set predetermined parameter q to the true maximum indegree of a graph.

For the US algorithm, Fisher's independence test was exploited with the significance level $\alpha = 1 - \Phi(0.5n^{1/3})$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. For the GDS algorithm, we set the initial graph to a random graph. Since the GDS algorithm uses a greedy search, its accuracy relies on the initial graph, so the GDS algorithm can recover the graph better with an appropriate choice of an initial graph. Finally, the US and GDS algorithms were not applied to high-dimensional models ($n < p$) because they are designed only for low-dimensional linear SEMs.

4.1 Gaussian Linear SEMs

We first conducted simulations using 100 realizations of p-node Gaussian linear SEMs (4) with randomly generated underlying DAG structures for node size $p \in \{25, 50, 100, 150, 200, 250\}$ while respecting the maximum degree constraint $d \in \{5, 8\}$ as done by Ghoshal and Honorio (2017). The set of non-zero parameters, $\beta_{jk} \in \mathbb{R}$ in Equation (4), was generated uniformly at random in the range $\beta_{jk} \in (-0.6, -0.4) \cup (0.4, 0.6)$. Lastly, all noise variances were set to $\sigma_j^2 = 0.75$.

Figures 1 (a) and (c) show the empirical probability of successful DAG recovery with Algorithm 1 by varying sample size $n \in \{100, 200, ..., 1500\}$ for $d = 5$ and $n \in \{100, 200, ..., 2000\}$



(a) Sparse: $d = 5$

(b) Sparse: $d = 5$

(c) Dense: $d = 8$

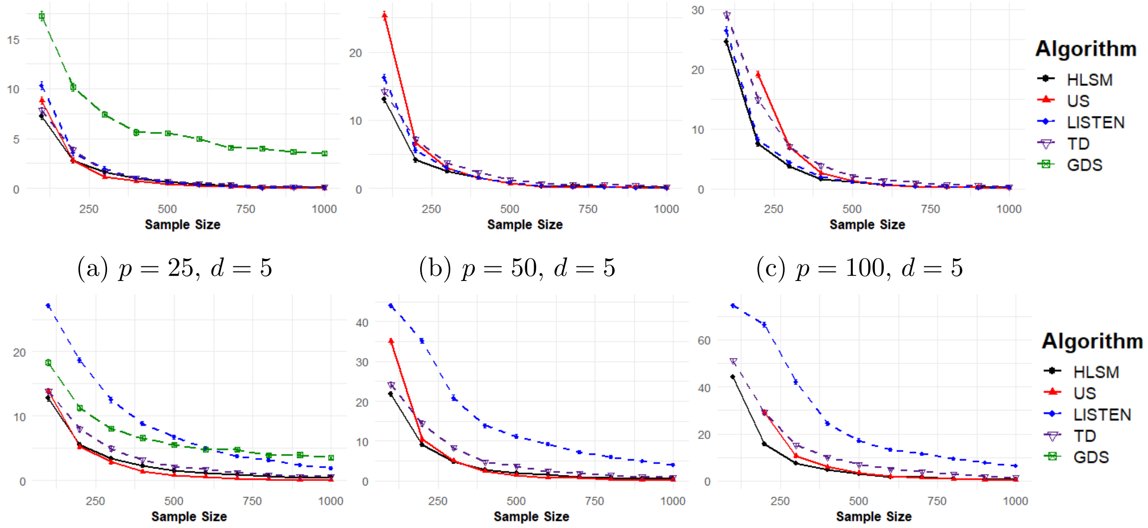
(d) Dense: $d = 8$

Figure 1: Probabilities of successful structure recovery for Gaussian linear SEMs with maximum degree $d \in \{5, 8\}$. The empirical probability of successful directed graph recovery is shown versus sample size n (left) and versus re-scaled sample size $C = n/(\log p)$ (right).

for $d = 8$, respectively. Figures 1 (b) and (d) plot the empirical probability against re-scaled sample size $C = n/\log p$. This confirms Corollary 9 that sample size n required for a successful graph structure recovery scales logarithmically with the number of nodes p. Hence, we would expect the empirical curves for different problem sizes to more closely align with this re-scaled sample size on the horizontal axis, a result clearly seen in Figures 1 (b) and (d). Figures 1 (a) - (d) also reveal that Algorithm 1 requires fewer samples to recover a sparse graph. Hence, these simulation results empirically support our theoretical findings.

Figure 2 evaluates the proposed algorithm and the state-of-the-art US, GDS, LISTEN, and TD algorithms in terms of recovering DAGs with $p \in \{25, 50, 100\}$ by varying sample size $n \in \{100, 200, ..., 1000\}$. As seen in Figure 2, the proposed algorithm (HLSM) recovers the true directed edges better as the sample size increases and the Hamming distance converges to 0. We also see that the proposed algorithm performs better for the sparse setting ($d = 5$) than for the dense setting ($d = 8$). Hence, these simulation results also heuristically confirm the consistency of the proposed algorithm.

Figure 2 shows that the proposed algorithm generally performs as accurately as, or significantly better than, the comparison algorithms with our settings. This phenomenon is not contradictory, because the comparison methods are designed for learning (sub-)Gaussian



(a) $p = 25$, $d = 5$ (b) $p = 50$, $d = 5$ (c) $p = 100$, $d = 5$

(d) $p = 25, d = 8$ (e) $p = 50, d = 8$ (f) $p = 100, d = 8$

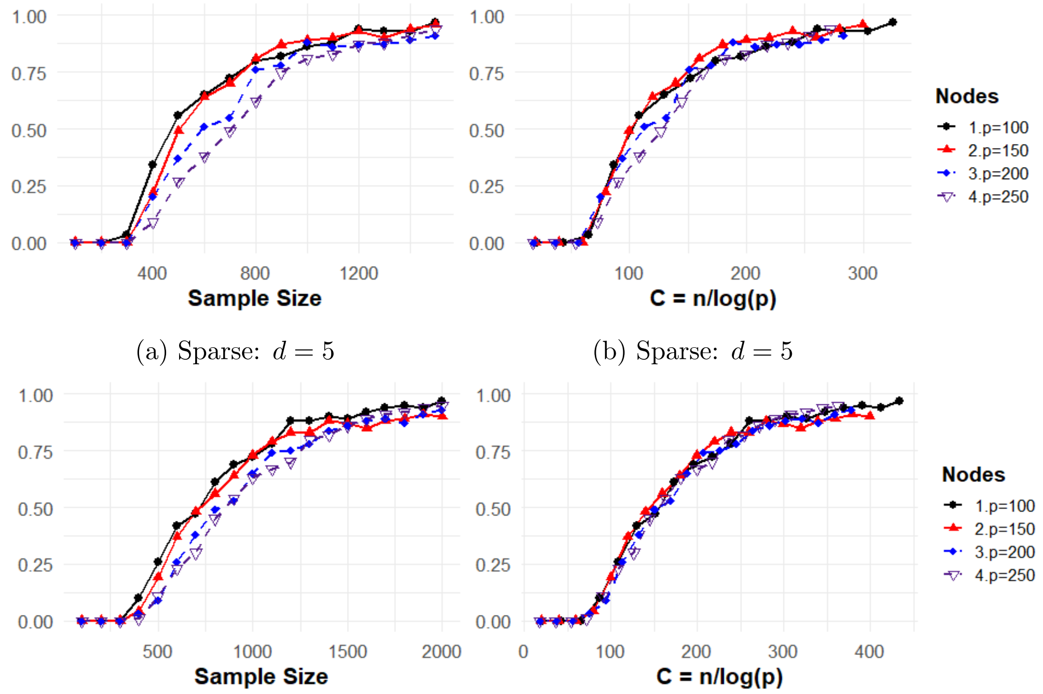
Figure 2: Comparison of the proposed algorithm (HLSM) against the US, GDS, LISTEN, and TD algorithms in terms of average Hamming distance for learning Gaussian linear SEMs.

linear SEMs in low- and high-dimensional settings. However, it must be emphasized that the comparison US and GDS algorithms cannot be implemented in high-dimensional or large-scale graph settings. Furthermore, the TD algorithm is not applicable to large-scale dense graphs owing to the heavy computational cost discussed in Section 3.2. Hence, we do not present the results of the comparison methods for large-scale graphs because of the lack of samples and the huge run time, but a comparison of the run times is provided in Section 4.3.

Lastly, we again point out that the performances of the US, GDS, and LISTEN algorithms depend highly on the significance level, the initial graph, and the regularization parameter, respectively. In addition, unlike the proposed and LISTEN algorithms, the sample complexity of the TD algorithm relies on the maximum indegree of a true graph rather than the maximum degree of the moralized graph. Hence, we emphasize that this numerical study does not imply that the proposed method is always better than the comparison algorithm.

4.2 Linear SEMs with Different Error Distributions

This section verifies the main result that Algorithm 1 successfully learns high-dimensional linear SEMs where non-Gaussian error distributions are allowed. Hence, 100 sets of samples were generated under the procedure specified in Sections 4.1, except that error distributions were sequentially Uniform, $U(-1.25, 1.25)$, Gaussian, $N(0, 0.5)$, a $\frac{1}{\sqrt{3}}$ Student's t with 6 degree of freedom, and a twice centered Beta, $Beta(0.5, 0.5)$, distributions. Then, the



(a) Sparse: $d = 5$

(b) Sparse: $d = 5$

(c) Dense $d = 8$

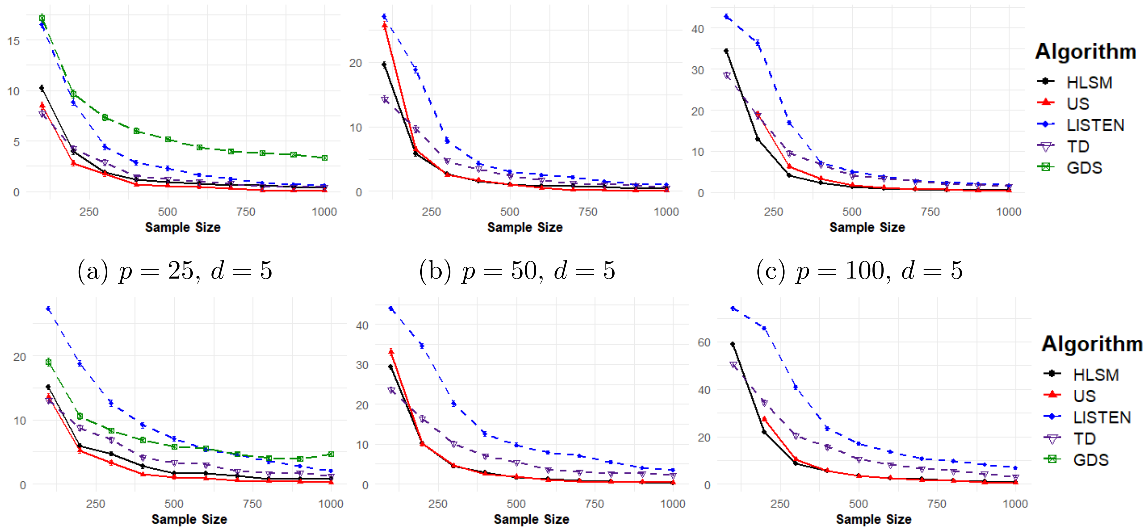
(d) Dense: $d = 8$

Figure 3: Probabilities of successful structure recovery for linear SEMs with maximum degree $d \in \{5, 8\}$. The empirical probability of successful directed graph recovery is shown versus the sample size n (left) and versus re-scaled sample size $C = n/(\log p)$ (right).

proposed algorithm and the comparison methods were evaluated by varying the sample size, as seen in Figures 3 and 4.

The simulation results in Figure 3 and 4 are analogous to the results for Gaussian linear SEMs in Section 4.1. More specifically, they empirically support the assertion that the proposed algorithm requires sample size n depending on maximum degree d and $\log p$ for successful graph structure recovery. Hence, these numerical experiments confirm that, under the required conditions, Algorithm 1 consistently learns high-dimensional sparse linear SEMs, regardless of the type of error distribution.

Further shown is that the proposed algorithm at our settings recovers the graph as accurately as, and better than, the comparison algorithms in terms of Hamming distance. That is also expected, because Peters and Bühlmann (2014); Chen et al. (2019) empirically show that the GDS and TD algorithms can successfully learn linear SEMs even with heterogeneous error variances. Furthermore, Park (2020) discusses the robustness of the US algorithm with non-Gaussian linear SEMs without theoretical guarantees. However, the GDS algorithm is only applied to 25-node graphs because of the huge computational cost.



(a) $p = 25,\ d = 5$ (b) $p = 50,\ d = 5$ (c) $p = 100,\ d = 5$

(d) $p = 25, d = 8$ (e) $p = 50, d = 8$ (f) $p = 100, d = 8$

Figure 4: Comparison of the proposed algorithm (HLSM), against the US, GDS, LISTEN, and TD algorithms in terms of average Hamming distance when learning linear SEMs with Gaussian, Uniform, Student's t error, and Beta error distributions.

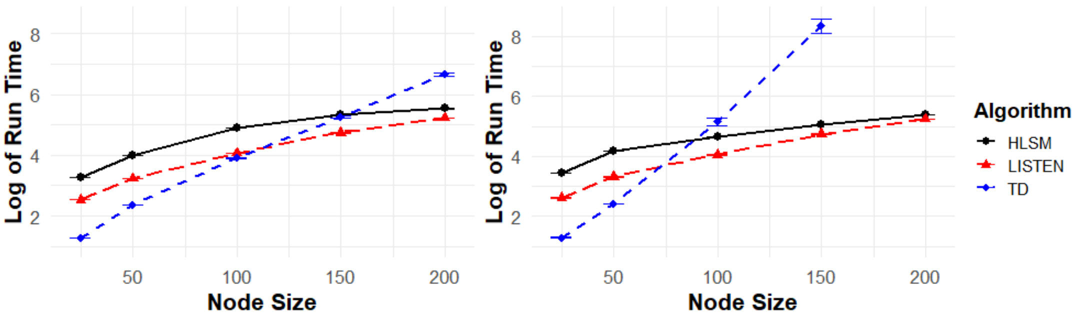
4.3 Computational Complexity

One of the important issues in learning DAG models is computational cost owing to the super-exponentially growing number of DAGs in the number of nodes. To validate the computational complexity discussed in Section 3.2, Figure 5 compares the average log run time of high-dimensional linear SEM learning algorithms. More specifically, Figure 5 measured the run time of the proposed, LISTEN and TD algorithms when learning Gaussian SEMs, exploited in Section 4.1, with fixed sample size $n = 1000$ by varying node size $p \in \{25, 50, 100, 150, 200\}$ and maximum degree $d \in \{5, 8\}$.

As we can see in Figure 5, the proposed algorithm has polynomial computational complexity in the number of nodes. In addition, as the number of nodes increases, the proposed algorithm is computationally more efficient than the comparison methods. This phenomenon is more exaggerated in our settings when the considered graphs are dense. For a 200-node dense graph estimation especially, the average run time of the TD algorithm takes over 10 hours, and hence, its result does not appear in Figure 5. However, we again point out that the computational complexity of the TD algorithm depends on the maximum indegree of a graph. Hence, when the maximum indegree of a true graph is small, while the degree of its moralized graph is large, the TD algorithm is computationally more efficient.

5. Real Data: Spread Map of COVID-19 in China

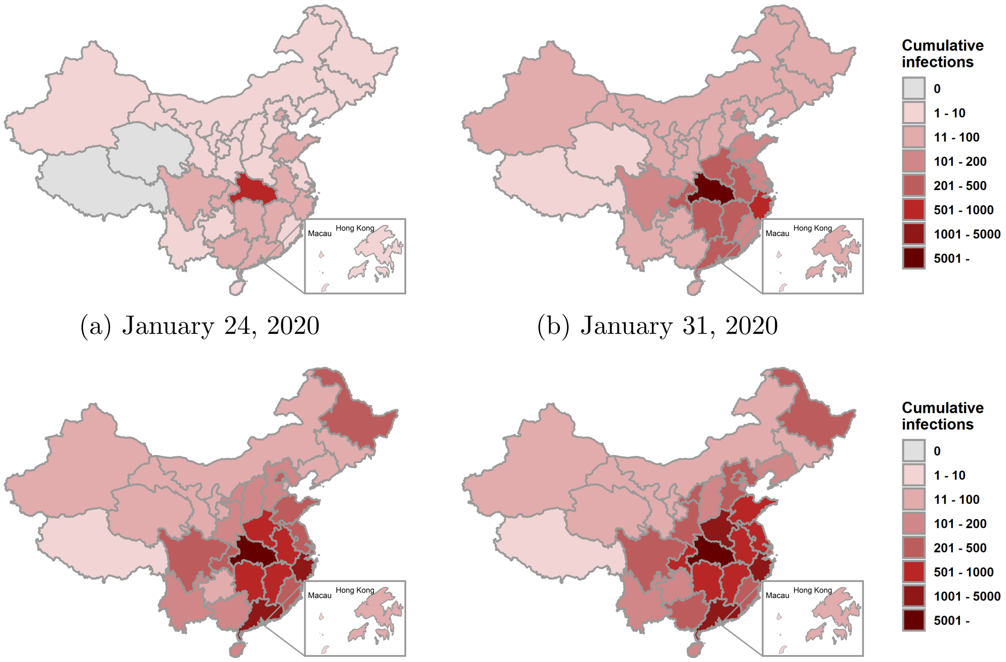
This section applies the proposed algorithm to real COVID-19 data for daily confirmed cases in China where the COVID-19 viral disease first spread in Wuhan and became highly



(a) Sparse: $d = 5$

(b) Dense: $d = 8$

Figure 5: Comparison of the proposed algorithm (HLSM), against the LISTEN and TD algorithms in terms of average log run time for learning Gaussian linear SEMs when sample size $n = 1000$ and number of nodes $p \in \{25, 50, 100, 150, 200\}$.



(a) January 24, 2020

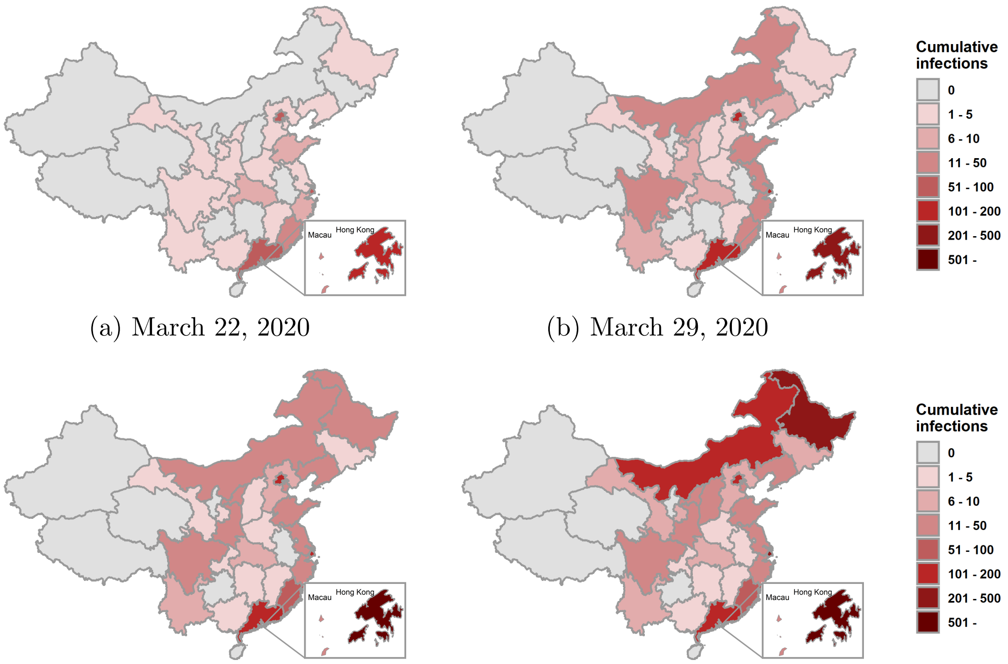
(b) January 31, 2020

(c) February 07, 2020

(d) February 14, 2020

Figure 6: Heat maps of the cumulative confirmed cases for major cities and provinces in China from January 24, 2020, to February 14, 2020.

contagious (Wu et al., 2020). Although the disease spread may not be acyclic, the violation of the assumptions was considered as additive errors. Hence, the estimated graph provides the major trend of the disease spread rather than the exact spread. The data were



(a) March 22, 2020

(b) March 29, 2020

(c) April 5, 2020

(d) April 12, 2020

Figure 7: Heat maps of the cumulative confirmed cases for major cities and provinces in China from March 22, 2020, to April 12, 2020.

collected from the Coronavirus Resource Center of Johns Hopkins University & Medicine (https://coronavirus.jhu.edu/map.html).

We focused on 31 major cities and provinces and the daily percentage of new infections over each of the following two periods: Stage (1) from January 24, 2020 (the early days of the COVID-19) to March 15, 2020 (when transmission of the virus stabilized), and Stage (2) from March 16, 2020 (when the number of confirmed cases began to increase again) to May 6, 2020. Hence, the data set for each stage has a total of 52 records by date for the following 31 regions in China: Anhui, Beijing, Chongqing, Fujian, Gansu, Guangdong, Guangxi, Guizhou, Hainan, Hebei, Heilongjiang, Henan, Hong Kong, Hunan, Inner Mongolia, Jiangsu, Jiangxi, Jilin, Liaoning, Macau, Ningxia, Qinghai, Shaanxi, Shandong, Shanghai, Shanxi, Sichuan, Tianjin, Xinjiang, Yunnan, and Zhejiang. In summary, both data sets contain $p = 31$ covariates and $n = 52$ samples. However, Hubei and Tibet were excluded from the analysis because Wuhan, a major city in Hubei, had been locked down from January 23 to April 8, and Tibet had only one confirmed case (a person who had traveled to Wuhan). Lastly, owing to skewness in the data from outliers, we exploited a 3-day moving average for each day calculated by averaging the values of that day, the day before, and the next day.

Figures 6 and 7 show the heat maps for the total number of confirmed cases for Stages (1) and (2), respectively. However, for Figure 7, the confirmed cases before March 15 were excluded to focus only on how the coronavirus spread in Stage (2). Hence, they uncover the approximate infection status, and how COVID-19 spread in Stages (1) and (2). As can be seen in Figure 6, comparatively, there were a lot of confirmed cases in south-central and eastern China, in places such as Henan, Guangdong, and Zhejiang, whereas northern and north-western China had only a few confirmed cases on January 24. However, at the end of Stage (1) on February 14, a lot of confirmed cases were seen in all provinces. Hence, it seemed that COVID-19 was spreading outward from south-central China.

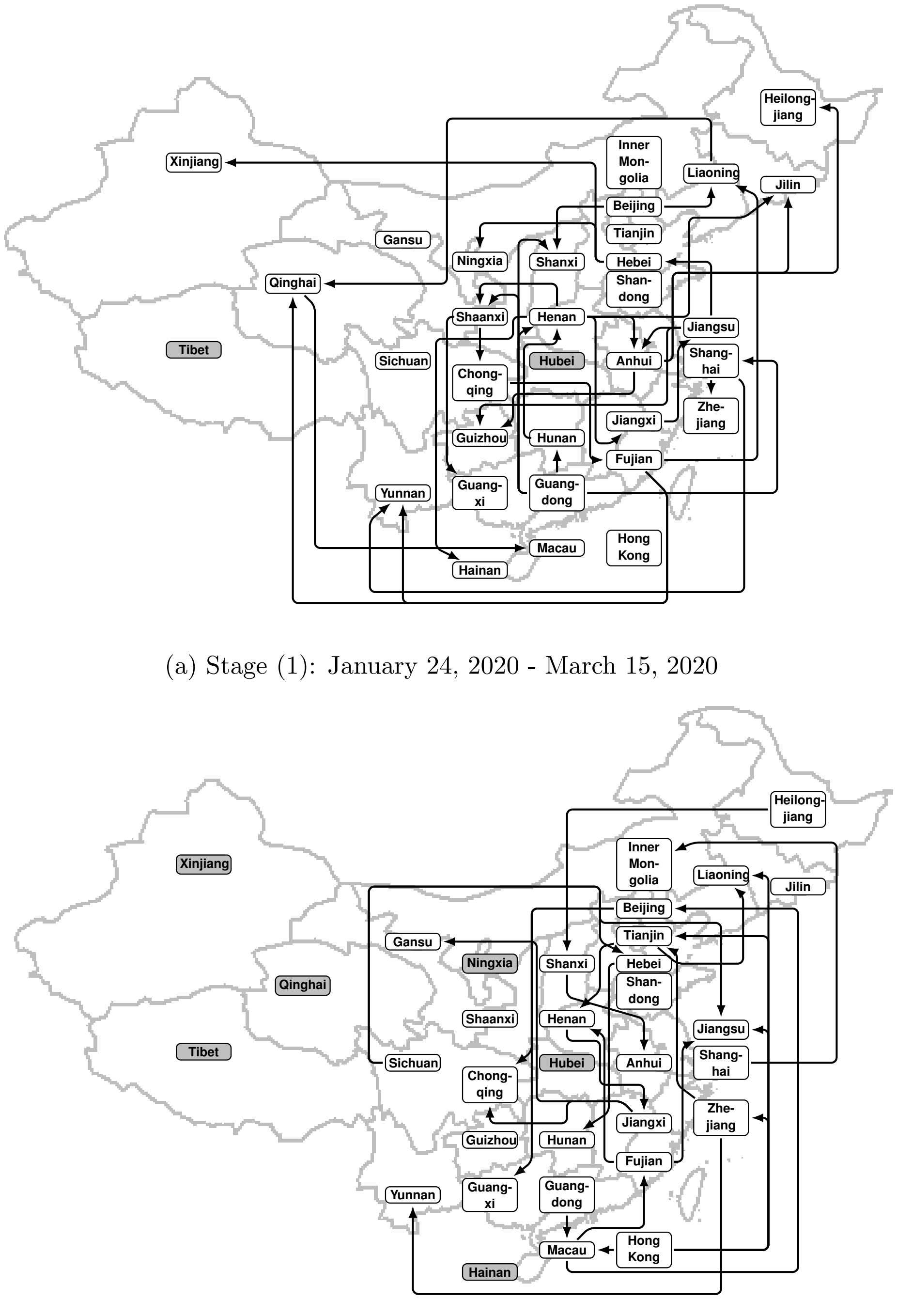
Figure 7 shows a lot of confirmed cases in southern China (for example, in Hong Kong) on March 16. However, at the end of Stage (2), on April 12, there were more confirmed cases reported in northern and north-eastern China. This phenomenon might be from a spread of infections from Hong Kong to northern and north-eastern China in Stage (2). In the same manner, it seems the disease spread from Beijing to adjacent provinces. Lastly, there were no confirmed cases were reported in Hainan, Ningxia, Qinghai, and Xinjiang in Stage (2). Hence, they were not considered for the analysis of COVID-19 spread in Stage (2).

Figures 8 (a) and (b) show directed graphs estimated by the proposed algorithm for Stages (1) and (2), respectively, where the regularization parameter was set to $\lambda = 7.5 \times \sqrt{\frac{\log p}{n}}$ in order to see only legitimate edges. As we can see in Figure 8 (a), there were 32 directed edges, and most of the directed edges were from south-central China toward other regions. In particular, there were 10 edges from Guangdong and Henan, whereas the regions in north-eastern and north-western provinces, such as Xinjiang, Jilin, and Heilongjiang, had only incoming edges. We believe this result agrees with Figure 6, which shows the disease movement trend from south-central China toward other districts. Hence, it seems to make sense for there to be many directed edges from Guangdong and Henan. In the same manner, most of the estimated directed edges are interpretable.

However, we also acknowledge that there were some non-explainable edges between spatially far-away districts, such as (Liaoning, Qinghai), (Fujian, Qinghai), and (Shanghai, Yunan). In addition, some important edges are missing between adjacent provinces, such as edges from Zhejiang, which was one of the most infectious areas in the early days of Stage (1). These falsely estimated and missing edges are also shown in the estimated graph for Stage (2). That might have been caused by the following: (i) the data used in this paper are not independent; (ii) the required conditions for the proposed algorithm may not be satisfied; and (iii) inner-province infections and mutual transmissions between provinces existed.

Figure 8 (b) presents 25 directed edges, and we can see a trend where most of the edges are toward northern and north-eastern China. Furthermore, there were 5 directed edges from Hong Kong, which is consistent with Figure 7, where there was a trend of infections spreading from Hong Kong to northern and north-eastern China.

Finally, we acknowledge that our proposed method has many errors, because other neighboring countries were not considered and our assumptions, such as the existence of clear cycles, nonlinear dependency and independent samples, may not be completely satisfied. Nevertheless, it should be emphasized that this analysis does not apply any prior information on COVID-19's spread, but only exploits the number of daily confirmed cases in China. Hence, we believe the proposed method is applicable as a low-resolution way to figure out how a viral disease spreads, even when there is no prior tracking information.



(a) Stage (1): January 24, 2020 - March 15, 2020

(b) Stage (2): March 16, 2020 - May 6, 2020

Figure 8: Estimated directed graphs for new COVID-19 confirmed-case proportions in China via the proposed algorithm.

6. Discussion and Future Works

This paper provides a statistically consistent and computational feasible algorithm for learning high-dimensional linear SEMs via ℓ_1-regularized regression. However, several topics remain for future work. An important problem is determining whether the backward stepwise selection identifiability condition is satisfied from the observational data. However, not yet studied is how it can be confirmed from the data. The proposed method requires a sparse moralized graph that could be restrictive for special types of graphs, such as a bipartite graph and a star graph. Hence, it would be interesting to explore a new approach with relaxed sparsity condition (e.g., a maximum indegree constraint or approximate sparsity condition in Klaassen et al., 2018). Furthermore, the proposed algorithm needs restrictive linearity, incoherence and minimum signal assumptions. We believe that a new method with milder conditions can be developed, and one may be able to prove its consistency.

In addition, we conjecture that the proposed algorithm can be improved in terms of the computational cost, because the algorithm may not require $O(p^2)$ ℓ_1-regularized regressions. Hence, we believe it is possible to develop a more computationally efficient algorithm. Lastly, we also believe our regression-based algorithm can easily be extended to robust linear SEM learning, like the undirected graphical model learning approaches.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2021R1C1C1004562)

Appendix A. Proof for Theorem 7

Proof Algorithm 1 involves $p(p-1)/2$ optimization problems in Equation (10). However, the theoretical guarantees for all problems are analogous, and hence, we only consider an arbitrary $r \in \{1, 2, ..., p-1\}$-th iteration, and node $j \in \{\pi_1, ..., \pi_{p+1-r}\}$ in Algorithm 1. For ease of notation, let $[\cdot]_k$ and $[\cdot]_S$ denote parameter(s) corresponding to variable X_k and random vector X_S, respectively. In addition, subscripted j and r in parentheses are omitted to improve readability, and hence, we denote $\theta_j^*(r)$, $\widehat{\theta}_j(r)$, $S_j(r)$, $T_j(r)$, $Z_j(r)$, and $\widehat{Z}_j(r)$ as θ^*, $\widehat{\theta}$, S, T, Z, and $\widehat{Z}$, respectively. Furthermore, we let $s_{\max}$ and $K_{\max}$ are the maximum values of the sub-Gaussian and the $(4m)$-th bounded moment parameters of variables, respectively. Lastly, $\sigma_{\max}^2$ is the maximum error variance.

We restate the true parameters in Equation (7). Suppose that θ^* denotes the solution to the following problem, where $S \subset V \setminus (\{j\} \cup \{\pi_p, ..., \pi_{p+2-r}\})$ if $r \geq 2$; otherwise, $S = V \setminus \{j\}$:

$$\theta^* := \arg\min_{\theta \in \mathbb{R}^{|S|}} \mathbb{E}\left((X_j - \langle X_S, \theta \rangle)^2 \right). \tag{8}$$

Simple algebra yields $\Sigma_{SS}\ \theta^* = \Sigma_{Sj}$ where Σ_{SS} is a sub-matrix of the true covariance matrix Σ corresponding to variables X_S. Hence, we re-formalize X_j as the following:

$$X_j = \langle X_{\mathrm{Pa}(j)}, \beta^* \rangle + \epsilon_j = \langle X_S, \theta^* \rangle + \delta_j,$$

where $\delta_j = X_j - \langle X_S, \theta^* \rangle$.

For ease of notation, we define a set of non-zero elements as an index of θ^*, which is $T := \{k \in S \mid [\theta^*]_k \neq 0\}$. Then, X_j can be re-written as

$$X_j = \langle X_S, \theta^* \rangle + \delta_j = \langle X_T, [\theta^*]_T \rangle + X_j - \langle X_T, [\theta^*]_T \rangle. \tag{9}$$

The main goal of the proof is to find a minimizer of the following ℓ_1-regularized regression problem:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{|S|}}{\text{argmin}} \; \frac{1}{2n} \sum_{i=1}^{n} \left(X_j^{(i)} - \langle X_S^{(i)}, \theta \rangle \right)^2 + \lambda \|\theta\|_1, \tag{10}$$

where $\lambda > 0$ is a regularization parameter. By setting the *sub-differential* to 0, $\widehat{\theta}$ satisfies the following condition:

$$\nabla_\theta \left(\frac{1}{2n} \sum_{i=1}^{n} (X_j^{(i)} - \langle X_S^{(i)}, \theta \rangle)^2 + \lambda \|\theta\|_1 \right) = \frac{1}{n} \sum_{i=1}^{n} \left(X_j^{(i)} - \langle X_S^{(i)}, \theta \rangle \right) X_S^{(i)} + \lambda \widehat{Z} = 0, \tag{11}$$

where $\widehat{Z} \in \mathbb{R}^{|S|}$ and $[\widehat{Z}]_t = \text{sign}([\widehat{\theta}]_t)$ if $t \in T$; otherwise, $|[\widehat{Z}]_t| \leq 1$.

In the high-dimensional setting ($p > n$), the convex program in Equation (10) is not necessarily strictly convex, so it might have multiple optimal solutions. Hence, we introduce the following lemma, adapted from Lemma 1 in Ravikumar et al., 2010 and Lemma 8 in Yang et al., 2015, implying that the solutions nonetheless share their support set under appropriate conditions.

Lemma 10 (Lemma1 in Ravikumar et al., 2010, Lemma 8 in Yang et al., 2015) *For any $r \in \{1, 2, ..., p-1\}$-th iteration, and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, suppose that $|[\widehat{Z}_j(r)]_t| < 1$ for $t \notin T_j(r)$ in Equation (11). Then, any solution $\widehat{\theta}_j(r)$ of Equation (10) satisfies $[\widehat{\theta}_j(r)]_t = 0$ for all $t \notin T_j(r)$. Furthermore, if $\frac{1}{n} X_{T_j(r)}^\top X_{T_j(r)}$ is invertible, then $\widehat{\theta}$ is unique.*

Applying Assumption 4, Lemma 10 ensures that the solution to the ℓ_1-regularized regression is unique, as long as $|[\widehat{Z}]_t| < 1$ for all $t \notin T$. Hence, the remainder of the proof is to show $|[\widehat{Z}]_t| < 1$ for all $t \notin T$.

The following lemma provides the probability bound that $\max_{t \in T^c} |[\widehat{Z}]_t|$ is less than 1 for sub-Gaussian and bounded-moment error distributions, respectively.

Lemma 11 *Consider a fixed $r \in \{1, 2, ..., p-1\}$-th iteration, $j \in \{\pi_1, ..., \pi_{p+1-r}\}$ and $\lambda < \frac{10\sigma_{\max}^2}{\gamma}$.*

- *For a sub-Gaussian linear SEM,*

$$\Pr\left(\max_{t \in T_j^c(r)} |[\widehat{Z}_j(r)]_t| < 1 \right)$$
$$\geq 1 - 4 \cdot exp\left(\frac{-n}{128(1 + 4s_{\max}^2) \max_j (\Sigma_{jj})^2} \frac{\lambda^2 \gamma^2 \lambda_{\min}^4}{(d+2)^2 (10\sigma_{\max}^4 + \lambda\gamma\sigma_{\max}^2 + \lambda\gamma\lambda_{\min})^2} \right).$$

- *For a (4m)-th bounded-moment linear SEM,*

$$\Pr\left(\max_{t \in T_j^c(r)} |[\widehat{Z}_j(r)]_t| < 1\right) \\ \geq 1 - 4 \cdot \frac{2^{2m} \max_j(\Sigma_{jj})^{2m} C_m (K_{\max}+1)}{n^m} \left(\frac{\lambda^2 \gamma^2 \lambda_{\min}^4}{(d+2)^2 (10\sigma_{\max}^4 + \lambda\gamma\sigma_{\max}^2 + \lambda\gamma\lambda_{\min})^2}\right)^{-m},$$

where C_m is a constant depending only on m.

So far, we have shown that the solution $\widehat{\theta}$ to Equation (10) satisfies $[\widehat{\theta}]_t = 0$ for all $t \in T^c$ with high probability. Now, we focus on $\text{sign}([\widehat{\theta}]_t) = \text{sign}([\theta]_t)$ for all $t \in T$. The following lemma provides the maximum error bound of each component of $[\widehat{\theta}]_t$ for any $t \in T$.

Lemma 12 *Consider a fixed $r \in \{1, 2, ..., p-1\}$-th iteration, $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, and an arbitrary small positive $\epsilon' \in (0, \theta_{\min})$.*

- *For a sub-Gaussian linear SEM,*

$$\Pr\left(\max_{j \in V} \|[\widehat{\theta}_j]_{T_j(r)} - [\theta_j]^*_{T_j(r)}\|_\infty \leq \epsilon'\right) \\ \geq 1 - 4 \cdot exp\left(\frac{-n}{128(1+4s_{\max}^2)\max_j(\Sigma_{jj})^2}\left(\frac{1}{C_1|T_j(r)|}\left(\epsilon' - \frac{\lambda}{\lambda_{\min}}\right)\right)^2\right).$$

- *For a (4m)-th bounded-moment linear SEM,*

$$\Pr\left(\max_{j \in V} \|[\widehat{\theta}_j]_{T_j(r)} - [\theta_j]^*_{T_j(r)}\|_\infty \leq \epsilon'\right) \\ \geq 1 - 4 \cdot \frac{2^{2m} \max_j(\Sigma_{jj})^{2m} C_m (K_{\max}+1)}{n^m} \left(\frac{1}{C_1|T_j(r)|}\left(\epsilon' - \frac{\lambda}{\lambda_{\min}}\right)\right)^{-2m},$$

where C_m is a constant depending only on m.

Applying Assumption 6, $\min_{t \in T} |[\theta^*]_t| \geq \theta_{\min}$ and setting $\epsilon' = \theta_{\min}$, it is clear that $\text{sign}([\widehat{\theta}]_t) = \text{sign}([\theta^*]_t)$ for all $t \in T$ with high probability. Hence, combining the results shown above, we reach one of our main theoretical results. For a sub-Gaussian error linear SEM,

$$\Pr\left(\text{sign}\left([\widehat{\theta}]_{T_j(r)}\right) = \text{sign}\left([\theta^*]_{T_j(r)}\right)\right) \\ \geq 1 - 4 \cdot \exp\left(\frac{-n}{128(1+4s_{\max}^2)\max_j(\Sigma_{jj})^2}\left(\frac{1}{C_1|T_j(r)|}\left(\theta_{\min} - \frac{\lambda}{\lambda_{\min}}\right)\right)^2\right).$$

In addition for a $(4m)$-th bounded-moment linear SEM,

$$\Pr\left(\text{sign}\left([\widehat{\theta}]_{T_j(r)}\right) = \text{sign}\left([\theta^*]_{T_j(r)}\right)\right) \\ \geq 1 - 4 \cdot \frac{2^{2m} \max_j(\Sigma_{jj})^{2m} C_m (K_{\max}+1)}{n^m} \left(\frac{1}{C_1|T_j(r)|}\left(\theta_{\min} - \frac{\lambda}{\lambda_{\min}}\right)\right)^{-2m}.$$

Therefore, for a sub-Gaussian linear SEM, there exist positive constants $C_1, C_2, C_3, C_4 > 0$ such that

$$\Pr\left(\text{sign}\left(\widehat{\theta}_j(r)\right) = \text{sign}\left(\theta_j^*(r)\right)\right) \geq 1 - C_1 \cdot \exp\left(\frac{-C_2 n \lambda^2}{d^2}\right) - C_3 \cdot \exp\left(\frac{-C_4 n}{d^2}\left(\theta_{\min} - \frac{\lambda}{\lambda_{\min}}\right)^2\right).$$

In addition for a $(4m)$-th bounded-moment linear SEM, there exist positive constants $D_1, D_2 > 0$ such that

$$\Pr\left(\text{sign}\left(\widehat{\theta}_j(r)\right) = \text{sign}\left(\theta_j^*(r)\right)\right) \geq 1 - D_1 \cdot \frac{d^{2m}}{n^m \lambda^{2m}} - D_2 \cdot \frac{d^{2m}}{n^m}\left(\theta_{\min} - \frac{\lambda}{\lambda_{\min}}\right)^{-2m}.$$

■

Appendix B. Proof for Theorem 8

Proof This section proves that Algorithm 1 accurately estimates the ordering with high probability, given that $\theta_j^*(r)$ are well estimated from ℓ_1-regularized regression. This theorem can be proved in the same manner as the one developed in Park (2020). Here, we restate the proof in our framework.

Without loss of generality, assume that the true ordering is unique, and that $\pi = (\pi_1, \pi_2, ..., \pi_p) = (1, 2, ..., p)$. In addition for ease of notation, let $\pi_{1:j} = (\pi_1, \pi_2, ..., \pi_j)$, and omit subscripted j and r in parentheses as done in Appendix A. In addition, let $s_{\max}$ and $K_{\max}$ are the maximum values of sub-Gaussian and $(4m)$-th bounded moment parameter of variables, respectively. Lastly, $\sigma^2_{\max}$ is the maximum error variance. Then, the probability that the ordering is correctly estimated from Algorithm 1 is

$$\Pr\left(\widehat{\pi} = \pi\right) = \Pr\Big(\min_{\substack{j=2,...,p\\k=1,...,j-1}} \widehat{\text{Var}}(X_j \mid X_{\pi_{1:j}\setminus j}) - \widehat{\text{Var}}(X_k \mid X_{\pi_{1:j}\setminus k}) > 0\Big).$$

Since it can be decomposed into the following two terms, we have

$$\begin{aligned}\Pr\left(\widehat{\pi} = \pi\right) \geq \Pr\Big(&\min_{\substack{j=2,...,p\\k=1,...,j-1}} \Big\{\text{Var}(X_j \mid X_{\pi_{1:j}\setminus j}) - \text{Var}(X_k \mid X_{\pi_{1:j}\setminus k})\Big\} > \tau_{\min}, \text{ and}\\ &\max_{\substack{j=2,...,p\\k=1,...,j}} \Big|\text{Var}(X_k \mid X_{\pi_{1:j}\setminus k}) - \widehat{\text{Var}}(X_k \mid X_{\pi_{1:j}\setminus k})\Big| < \frac{\tau_{\min}}{2}\Big).\end{aligned}$$

The first term in the above probability is always satisfied because $\min\{\text{Var}(X_j \mid X_{\pi_{1:j}\setminus j}) - \text{Var}(X_k \mid X_{\pi_{1:j}\setminus k})\} > \tau_{\min}$ from Assumption 5. Hence, the probability that the ordering is correctly estimated from Algorithm 1 is reduced to

$$\Pr\left(\widehat{\pi} = \pi\right) \geq \Pr\left(\max_{\substack{j=2,...,p\\k=1,...,j}} \Big|\text{Var}(X_k \mid X_{\pi_{1:j}\setminus k}) - \widehat{\text{Var}}(X_k \mid X_{\pi_{1:j}\setminus k})\Big| < \frac{\tau_{\min}}{2}\right).$$

Applying the union bound, we have

$$\Pr\left(\widehat{\pi} = \pi\right) \geq 1 - p^2 \max_{j,k} \Pr\left(\left| \mathrm{Var}(X_k \mid X_{\pi_{1:j} \setminus k}) - \widehat{\mathrm{Var}}(X_k \mid X_{\pi_{1:j} \setminus k}) \right| < \frac{\tau_{\min}}{2} \right).$$

Now, we focus on the consistency rate of the two-stage ℓ_1-regularization based conditional variance estimator in Equation (6), which can be written as follows:

$$\widehat{\mathrm{Var}}(X_j \mid X_S) := \frac{1}{n} \sum_{i=1}^{n} \left(X_j - \langle X_T, \widehat{\alpha} \rangle \right)^2, \quad \text{where} \quad \widehat{\alpha} := \arg\min_{\alpha \in \mathbb{R}^{|T|}} \frac{1}{n} \sum_{i=1}^{n} \left(X_j^{(i)} - \langle X_T^{(i)}, \alpha \rangle \right)^2,$$

where $T := \{k \in S \mid [\theta^*]_k \neq 0\}$ as defined in Appendix A. We acknowledge that the above two-stage conditional variance estimator is well-known to be consistent. However, we prove its consistency in our settings where variables have the sub-Gaussian and bounded-moment properties.

Lemma 13 *Consider a fixed $r \in \{1, 2, ..., p-1\}$-th iteration, $j \in \{\pi_1, ..., \pi_{p+1-r}\}$.*

- *For a sub-Gaussian linear SEM,*

$$\Pr\left(\left| \widehat{Var}(X_j \mid X_{S_j(r)}) - Var(X_j \mid X_{S_j(r)}) \right| < \epsilon \right)$$
$$\geq 1 - 4 \cdot exp\left(\frac{-n}{128(1 + 4s_{\max}^2) \max_j (\Sigma_{jj})^2} \frac{\epsilon^2 \lambda_{\min}^4}{(d+1)^2 (\sigma_{\max}^2 (\epsilon + 5\sigma_{\max}^2) + \epsilon \lambda_{\min})^2} \right).$$

- *For a (4m)-th bounded-moment linear SEM,*

$$\Pr\left(\left| \widehat{Var}(X_j \mid X_{S_j(r)}) - Var(X_j \mid X_{S_j(r)}) \right| < \epsilon \right)$$
$$\geq 1 - 4 \cdot \frac{2^{2m} \max_j (\Sigma_{jj})^{2m} C_m (K_{\max} + 1)}{n^m} \left(\frac{\epsilon^2 \lambda_{\min}^4}{(d+1)^2 (\sigma_{\max}^2 (\epsilon + 5\sigma_{\max}^2) + \epsilon \lambda_{\min})^2} \right)^{-m},$$

where C_m is a constant depending only on m.

Hence, we complete the proof. For a sub-Gaussian linear SEM, there exist positive constants $C_1, C_2 > 0$ such that

$$\Pr\left(\widehat{\pi} = \pi\right) > 1 - C_1 p^2 \cdot \exp\left(\frac{-C_2 n}{(d+1)^2} \right).$$

For a $(4m)$-th bounded-moment linear SEM, there exists positive constant $D_1 > 0$ such that

$$\Pr\left(\widehat{\pi} = \pi\right) > 1 - D_1 p^2 \cdot \frac{(d+1)^{2m}}{n^m}.$$

■

Appendix C. Useful Propositions and their Proofs

C.1 Proof for Proposition 2

Proposition 2 For any $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, the true support of the solution $\theta_j^*(r)$ to Equation (7) is a subset of the neighborhood of j in the moralized graph:

$$\mathrm{Supp}(\theta_j^*(r)) \subset \mathrm{Ne}(j).$$

In addition, for $j = \pi_{p+1-r}$:

$$\mathrm{Supp}(\theta_j^*(r)) = \mathrm{Pa}(j).$$

Proof Without loss of generality, suppose $\Sigma_\epsilon = I_p$. For an arbitrary r-th iteration, consider a fixed node, $j \in \{\pi_1, ..., \pi_{p+1-r}\}$. Then, by definition, $S_j(r) = V \setminus (\{j\} \cup \{\pi_{p+1}, ..., \pi_{p+2-r}\})$ where $\pi_{p+1} = \emptyset$. For notional convenience, let $S = S_j(r)$ and $\pi = (1, 2, ..., p)$.

Suppose that Ω is the inverse covariance matrix for (X_j, X_S), which can be partitioned into four blocks using the Shur complement:

$$\Omega = \begin{bmatrix} \Omega_{jj} & \Omega_{jS} \\ \Omega_{jS}^\top & \Omega_{SS} \end{bmatrix} = \begin{bmatrix} (\Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj})^{-1} & -\Sigma_{jS}\Sigma_{SS}^{-1}(\Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj})^{-1} \\ -(\Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj})^{-T}\Sigma_{SS}^{-T}\Sigma_{Sj} & \Omega_{SS} \end{bmatrix},$$

where $\Sigma_{S_1 S_2}$ and $\Omega_{S_1 S_2}$ are sub-matrix of Σ and Ω corresponding to variables X_{S_1} and X_{S_2}, respectively. From the definition of $\theta_j^*(r)$ in Equation (7), we have

$$\theta_j^*(r) = -\frac{\Omega_{jS}}{\Omega_{jj}}.$$

In addition, from the definition of the linear SEM in Equation (2), for $k \in S$,

$$\Omega_{jk} = [I - B - B^\top + B^\top B]_{jk} = -\beta_{jk} - \beta_{kj} + \sum_{\ell \in S \setminus \{k\}} \beta_{\ell j}\beta_{\ell k}.$$

Note that for all $k \notin \mathrm{Ne}(j)$, $\beta_{jk} = \beta_{kj} = 0$, and $\beta_{\ell j}\beta_{\ell k} = 0$ for all $\ell \in V \setminus \{j, k\}$. Hence, we complete the proof that $[\theta_j^*(r)]_k = 0$ and $\mathrm{Supp}(\theta_j^*(r)) \subset \mathrm{Ne}(j)$.

Finally, for the special case $j = \pi_{p+1-r}$, S does not contain any descendant; that is, for all $k \in S$, $\beta_{kj} = 0$, and $\beta_{\ell j}\beta_{\ell k} = 0$ for all $\ell \in S \setminus \{k\}$. Hence, we obtain $-\Omega_{jk}$ an edge weight from k to j; that is, β_{jk}. Since there is no path cancellation involved, we obtain the final result, $\mathrm{Supp}(\theta_j^*(r)) = \mathrm{Pa}(j)$, without the faithfulness assumption. ■

C.2 Proof for Proposition 14

Proposition 14 *For any $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, the conditional variance of X_j given $S_j(r)$ satisfies the following.*

$$Var(X_j \mid X_{S_j(r)}) = Var(X_j \mid X_{T_j(r)}),$$

where $T_j(r)$ is support of the solution $\theta_j^*(r)$ in Equation (7).

Proof For an arbitrary r-th iteration, consider a fixed node $j \in V \setminus \{\pi_{p+1}, ..., \pi_{p+1-r}\}$. For notional convenience, let $S = S_j(r)$, $T = T_j(r)$, and $\theta^* = \theta_j^*(r)$.

Applying the properties of Equation (7) and the conditional expectation, we have

$$\mathbb{E}(X_j \mid X_S) = \mathbb{E}(\mathbb{E}(X_j \mid X_T) \mid X_S) = \mathbb{E}(\mathbb{E}(\langle X_T, [\theta^*]_T\rangle \mid X_T) \mid X_S) = \langle X_T, [\theta^*]_T\rangle.$$

Applying the law of total variance, $\mathrm{Var}(X_j \mid X_T)$ can be decomposed into

$$\begin{aligned}\mathrm{Var}(X_j \mid X_T) &= \mathbb{E}(\mathrm{Var}(X_j \mid X_S) \mid X_T) + \mathrm{Var}(\mathbb{E}(X_j \mid X_S) \mid X_T)\\ &= \mathbb{E}(\mathrm{Var}(X_j \mid X_S) \mid X_T) + \mathrm{Var}(\langle X_T, [\theta^*]_T\rangle \mid X_T)\\ &= \mathbb{E}(\mathrm{Var}(X_j \mid X_S) \mid X_T).\end{aligned}$$

Applying the Schur complement, we can see that there exists a function h such that

$$\begin{aligned}\mathrm{Var}(X_j \mid X_S) &= \Sigma_{jj} - \Sigma_{jS}\Sigma_{SS}^{-1}\Sigma_{Sj} = [\Sigma_{j\cup S, j\cup S}]_{jj}^{-1}\\ &= h\left((\beta_{jk})_{(k,j)\in E}, \sigma_{j\in V}^2\right).\end{aligned}$$

This result shows that $\mathrm{Var}(X_j \mid X_S)$ is not a function of random variables X. Hence, this completes the proof that

$$\mathrm{Var}(X_j \mid X_S) = \mathrm{Var}(X_j \mid X_T).$$

■

Appendix D. Useful Lemmas and their Proofs

D.1 Proof for Lemma 10

Lemma 10 (Lemma1 in Ravikumar et al., 2010, Lemma 8 in Yang et al., 2015) For any $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, suppose that $|[\widehat{Z}_j(r)]_t| < 1$ for $t \notin T_j(r)$ in Equation (11). Then, the solution $\widehat{\theta}_j(r)$ of Equation (10) satisfies $[\widehat{\theta}_j(r)]_t = 0$ for all $t \notin T_j(r)$. Furthermore, if $\frac{1}{n}X_{T_j(r)}^\top X_{T_j(r)}$ is invertible, then $\widehat{\theta}_j(r)$ is unique.

Proof This lemma can be proven in the same manner developed for special cases (Wainwright et al., 2006; Ravikumar et al., 2010). In addition, this proof is directly from Lemma 8 in Yang et al. (2015). Here, we restate the proof in our framework. For ease of notation, we again omit subscripted j and r in parentheses as in Appendix A.

In the main problem (10), all solutions have the same fitted value. That is mainly because if $\langle X, \theta_1\rangle \neq \langle X, \theta_2\rangle$ for some solutions θ_1 and θ_2, we have the following contradiction:

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^{n}\left(X_j^{(i)} - \delta\langle X_S^{(i)}, \theta_1\rangle - (1-\delta)\langle X_S^{(i)}, \theta_2\rangle\right)^2 + \lambda\|\delta\theta_1 + (1-\delta)\theta_2\|_1\\ &< \frac{\delta}{2n}\sum_{i=1}^{n}\left(X_j^{(i)} - \langle X_S^{(i)}, \theta_1\rangle\right)^2 + \frac{1-\delta}{2n}\sum_{i=1}^{n}\left(X_j^{(i)} - \langle X_S^{(i)}, \theta_2\rangle\right)^2 + \lambda\left(\|\delta\theta_1\|_1 + \|(1-\delta)\theta_2\|_1\right)\\ &= \min_{\theta}\left[\frac{1}{2n}\sum_{i=1}^{n}\left(X_j^{(i)} - \langle X_S^{(i)}, \theta\rangle\right)^2 + \lambda\|\theta\|_1\right],\end{aligned}$$

for any $\delta \in (0,1)$

Since any two solutions should have the same fitted value, they have the same squared error loss. Since all solutions should also achieve the same minimum value in the lasso problem (10), they must have the same ℓ_1-norm, $\|\theta_1\|_1 = \|\theta_2\|_1$. Furthermore, all solutions should satisfy the stationary condition $X_S^\top(X_j - \langle X_S, \widehat{\theta}\rangle) = \lambda \widehat{Z}$. Combining this with the uniqueness of fitted values, we conclude that sub-gradient $\widehat{Z}$ is unique for all solutions. The form of $\widehat{Z}$, where $[\widehat{Z}]_t = \text{sign}([\widehat{\theta}]_t)$ if $[\widehat{\theta}]_t \neq 0$, implies that $\langle \widehat{Z}, \widehat{\theta}\rangle = \|\widehat{\theta}\|_1$. Hence, for any other solutions $\widetilde{\theta}$, we have $\langle \widehat{Z}, \widetilde{\theta}\rangle = \|\widetilde{\theta}\|_1$. Briefly, using complementary slackness, we have $\langle \widehat{Z}, \widetilde{\theta}\rangle = \|\widetilde{\theta}\|_1$. This implies that for all index t for which $|[\widehat{Z}]_t| < 1$, $[\widetilde{\theta}]_t = 0$ (see Lemma1 of Ravikumar et al., 2010 for details). Therefore, if there exists a primal optimal solution $\widehat{\theta}$ with associated sub-gradient $\widehat{Z}$ such that $|[\widehat{Z}]_t| < 1$, then any optimal solution $\widetilde{\theta}$ also satisfies $[\widetilde{\theta}]_t = 0$ for all $t \in \{k \mid |[\widehat{Z}]_k| < 1\}$.

Finally, given that all optimal solutions satisfy $[\widetilde{\theta}]_t = 0$ for all $t \in T^c$, we may consider the restricted optimization problem subject to this set of constraints. Hence, if the principal sub-matrix of the Hessian is positive definite, $\frac{1}{n}X_T^\top X_T$, this sub-problem is strictly convex, so the optimal solution must be unique. ■

D.2 Proof for Lemma 11

Lemma 11 Consider a fixed $r \in \{1, 2, ..., p-1\}$-th iteration, $j \in \{\pi_1, ..., \pi_{p+1-r}\}$ and $\lambda < \frac{10\sigma_{\max}^2}{\gamma}$.

- For a sub-Gaussian linear SEM:

$$\Pr\left(\max_{t \in T_j^c(r)} |[\widehat{Z}_j(r)]_t| < 1\right)$$
$$\geq 1 - 4 \cdot \exp\left(\frac{-n}{128(1+4s_{\max}^2)\max_j(\Sigma_{jj})^2}\frac{\lambda^2\gamma^2\lambda_{\min}^4}{(d+2)^2(10\sigma_{\max}^4 + \lambda\gamma\sigma_{\max}^2 + \lambda\gamma\lambda_{\min})^2}\right).$$

- For a $(4m)$-th bounded-moment linear SEM:

$$\Pr\left(\max_{t \in T_j^c(r)} |[\widehat{Z}_j(r)]_t| < 1\right)$$
$$\geq 1 - 4 \cdot \frac{2^{2m}\max_j(\Sigma_{jj})^{2m} C_m (K_{\max}+1)}{n^m}\left(\frac{\lambda^2\gamma^2\lambda_{\min}^4}{(d+2)^2(10\sigma_{\max}^4 + \lambda\gamma\sigma_{\max}^2 + \lambda\gamma\lambda_{\min})^2}\right)^{-m},$$

where C_m is a constant depending only on m.

Proof The main idea of the proof is built on Lemma 5 in Harris and Drton (2013) and Lemmas 1 and 2 of Ravikumar et al. (2011) where the infinity norm of a matrix inversion error is considered when the sample covariance entries satisfy exponential-type tail and a polynomial-type tail bounds. Again, we restate the proof in our framework. For ease of notation, we again omit subscripted j and r in parentheses as in Appendix A. Therefore, $[\widehat{\theta}_j(r)]_T$, $[\widehat{Z}_j(r)]_T$, and $[\widehat{Z}_j(r)]_{T^c}$ are denoted by $\widehat{\theta}_T$, $\widehat{Z}_T$, and $\widehat{Z}_{T^c}$, respectively.

The idea of the primal-dual witness method is to plug in the sub-gradient of the estimated coefficients to true support. In order to obtain the closed form of $\widehat{Z}_{T^c}$, we divide Equation (11) into the following two parts:

$$\frac{1}{n}X_T^\top \left(X_j - \langle X_T, \widehat{\theta}_T \rangle\right) = \lambda \widehat{Z}_T, \quad \text{and} \quad \frac{1}{n}X_{T^c}^\top \left(X_j - \langle X_T, \widehat{\theta}_T \rangle\right) = \lambda \widehat{Z}_{T^c}. \tag{12}$$

Since Assumption 4 ensures $\frac{1}{n}X_T^\top X_T$ is invertible, Equation (12) derives the following:

$$\widehat{\theta}_T = \left(\frac{1}{n}X_T^\top X_T\right)^{-1}\left(\frac{1}{n}X_T^\top X_j - \lambda \widehat{Z}_T\right).$$

Applying $\widehat{\theta}_T$ to the second part of Equation (12), we have the following form of $\widehat{Z}_{T^c}$:

$$\begin{aligned}
\lambda \widehat{Z}_{T^c} &= \frac{1}{n}X_{T^c}^\top X_j - \frac{1}{n}X_{T^c}^\top X_T \left(\frac{1}{n}X_T^\top X_T\right)^{-1}\left(\frac{1}{n}X_T^\top X_j - \lambda \widehat{Z}_T\right) \\
&= \frac{1}{n}X_{T^c}^\top X_j - X_{T^c}^\top X_T \left(X_T^\top X_T\right)^{-1}\left(\frac{1}{n}X_T^\top X_j - \lambda \widehat{Z}_T\right) \\
&= \frac{1}{n}X_{T^c}^\top \left(I - X_T\left(X_T^\top X_T\right)^{-1}X_T^\top\right)X_j + \left(X_{T^c}^\top X_T\left(X_T^\top X_T\right)^{-1}\lambda \widehat{Z}_T\right).
\end{aligned}$$

Taking the ℓ_∞-norm of both sides, we have the following upper bound of $\|\widehat{Z}_{T^c}\|_\infty$:

$$\begin{aligned}
\|\widehat{Z}_{T^c}\|_\infty &\le \frac{1}{n\lambda}\|X_{T^c}^\top (I - P_T)X_j\|_\infty + \|X_{T^c}^\top X_T\left(X_T^\top X_T\right)^{-1}\widehat{Z}_T\|_\infty \\
&\le \frac{1}{n\lambda}\|X_{T^c}^\top (I - P_T)X_j\|_\infty + \left\|X_{T^c}^\top X_T\left(X_T^\top X_T\right)^{-1}\right\|_\infty,
\end{aligned}$$

where $P_T = X_T\left(X_T^\top X_T\right)^{-1}X_T^\top$.

Recalling the mutual incoherence assumption 3, $\|X_{T^c}^\top X_T\left(X_T^\top X_T\right)^{-1}\|_\infty \le (1-\gamma)$, we have

$$\|\widehat{Z}_{T^c}\|_\infty \le \frac{1}{n\lambda}\|X_{T^c}^\top (I - P_T)X_j\|_\infty + (1-\gamma).$$

This implies that $\|\widehat{Z}_{T^c}\|_\infty < 1$ if

$$\frac{1}{n}\|X_{T^c}^\top (I - P_T)X_j\|_\infty = \max_{k \in T^c}\left|\widehat{\Sigma}_{jk} - \widehat{\Sigma}_{jT}\widehat{\Sigma}_{TT}^{-1}\widehat{\Sigma}_{Tk}\right| < \frac{\lambda\gamma}{2},$$

where $\widehat{\Sigma}$ is the sample covariance matrix, and $\widehat{\Sigma}_{S_1 S_2}$ is its sub-matrix, corresponding to variables X_{S_1} and X_{S_2}.

From the definition of the partial correlation, we have

$$\rho_{j,k,T} = \frac{\Sigma_{jk} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tk}}{\sqrt{\Sigma_{jj} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tj}}\sqrt{\Sigma_{kk} - \Sigma_{kT}\Sigma_{TT}^{-1}\Sigma_{Tk}}},$$

The partial correlation is also obtained using the inversion of the covariance matrix,

$$\rho_{j,k,T} = -\frac{\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}},$$

where Ω is the inversion of the covariance matrix for (X_j, X_k, X_T).

Combining Proposition 2 and the facts that $\beta_{jk} = \beta_{kj} = 0$ and diagonal entries are $\Omega_{kk} = \frac{1}{\sigma_k^2} + \frac{1}{\sum_{\ell \in \mathrm{Ch}(k)\cap(T\cup j)} \beta_{k\ell}^2\sigma_\ell^2}$, we have

$$\Omega_{jj}^{-1} = \Sigma_{jj} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tj}\ , \Omega_{kk}^{-1} = \Sigma_{kk} - \Sigma_{kT}\Sigma_{TT}^{-1}\Sigma_{Tk}, \text{ and } \Sigma_{jk} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tk} = -\frac{\Omega_{jk}}{\Omega_{jj}\Omega_{kk}}.$$

Furthermore, from the conditional independence relationships, X_j and X_k are conditionally independent given X_T, and hence, $\Sigma_{jk} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tk} = \Omega_{jk} = 0$. Combining the results shown above, for $k \in T^c$, it is sufficient to show that

$$\begin{aligned}\left|\frac{1}{n}X_k^\top(I - P_T)X_j\right| &= \left|\left(\widehat{\Sigma}_{jk} - \widehat{\Sigma}_{jT}\widehat{\Sigma}_{TT}^{-1}\widehat{\Sigma}_{Tk}\right) - \left(\Sigma_{jk} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tk}\right)\right| \\ &= \left|\frac{\widehat{\Omega}_{jk}}{\widehat{\Omega}_{jj}\widehat{\Omega}_{kk}} - \frac{\Omega_{jk}}{\Omega_{jj}\Omega_{kk}}\right| < \frac{\lambda\gamma}{2}.\end{aligned}$$

Applying Lemma 16 and Lemma 17, the sufficient condition is as follows:

$$\|\widehat{\Omega} - \Omega\|_\infty \le \frac{\lambda\gamma}{\sigma_{\max}^2(\lambda\gamma + 10\sigma_{\max}^2)}.$$

Now, applying Lemma 15 (Lemma 5 in Harris and Drton, 2013), it is sufficient that $\widehat{\Sigma}$ satisfies

$$\|\widehat{\Sigma} - \Sigma\|_\infty < \frac{\lambda\gamma\lambda_{\min}^2}{(|T| + 2)(\sigma_{\max}^2(\lambda\gamma + 10\sigma_{\max}^2) + \lambda\gamma\lambda_{\min})}.$$

Applying the results of Lemma 18 (Lemmas 1 and 2 of Ravikumar et al., 2011), we complete the proof. For a sub-Gaussian linear SEM:

$$\begin{aligned}&\Pr\left(\max_{t\in T_j^c(r)} |[\widehat{Z}_j(r)]_t| < 1\right) \\ &\quad \ge 1 - 4\cdot\exp\left(\frac{-n}{128(1 + 4s_{\max}^2)\max_j(\Sigma_{jj})^2}\frac{\lambda^2\gamma^2\lambda_{\min}^4}{(d + 2)^2(\sigma_{\max}^2(\lambda\gamma + 10\sigma_{\max}^2) + \lambda\gamma\lambda_{\min})^2}\right).\end{aligned}$$

In addition for a $(4m)$-th bounded-moment linear SEM,

$$\begin{aligned}&\Pr\left(\max_{t\in T_j^c(r)} |[\widehat{Z}_j(r)]_t| < 1\right) \\ &\quad \ge 1 - 4\cdot\frac{2^{2m}\max_j(\Sigma_{jj})^{2m}C_m(K_{\max} + 1)}{n^m}\left(\frac{\lambda^2\gamma^2\lambda_{\min}^4}{(d + 2)^2(\sigma_{\max}^2(10\sigma_{\max}^2 + \lambda\gamma) + \lambda\gamma\lambda_{\min})^2}\right)^{-m},\end{aligned}$$

where C_m is a constant depending only on m. ■

D.3 Proof for Lemma 12

Lemma 12 Let $[\Delta]_{T_j(r)} := [\widehat{\theta}_j(r)]_{T_j(r)} - [\theta_j^*(r)]_{T_j(r)}$. For a sub-Gaussian linear SEM,

$$\Pr\left(\|[\Delta]_{T_j(r)}\|_\infty \le C_1|T_j(r)|\epsilon + \frac{\lambda}{\lambda_{\min}}\right) \ge 1 - 4\cdot\exp\left(\frac{-n\epsilon^2}{128(1+4s_{\max}^2)\max_j(\Sigma_{jj})^2}\right).$$

In addition, for a $(4m)$-th bounded-moment linear SEM,

$$\Pr\left(\|[\Delta]_{T_j(r)}\|_\infty \le C_1|T_j(r)|\epsilon + \frac{\lambda}{\lambda_{\min}}\right) \ge 1 - 4\cdot\frac{2^{2m}\max_j(\Sigma_{jj})^{2m}C_m(K_{\max}+1)}{n^m\epsilon^{2m}},$$

where C_m is a constant depending only on m.

Proof For better readability, we again omit subscripted j and r in parentheses. Therefore, $[\widehat{\theta}_j(r)]_T$ and $[\theta_j^*(r)]_T$ are denoted by $\widehat{\theta}_T$ and θ_T^*, respectively.

From the previous proof for Lemma 11, Equation (12) implies that

$$\begin{aligned}\widehat{\theta}_T &= \left(\frac{1}{n}X_T^\top X_T\right)^{-1}\left(\frac{1}{n}X_T^\top X_T\theta_T^* + \frac{1}{n}X_T^\top\delta_j - \lambda\text{sign}(\widehat{\theta}_T)\right)\\ &= \theta_T^* + \left(\frac{1}{n}X_T^\top X_T\right)^{-1}\left(\frac{1}{n}X_T^\top\delta_j - \lambda\text{sign}(\widehat{\theta}_T)\right). \qquad (13)\end{aligned}$$

Hence, we obtain the gap between $\widehat{\theta}_T$ and θ_T^*:

$$[\Delta]_T := \widehat{\theta}_T - \theta_T^* = \left(\frac{1}{n}X_T^\top X_T\right)^{-1}\left(\frac{1}{n}X_T^\top\delta_j - \lambda\text{sign}(\widehat{\theta}_T)\right).$$

Using the triangle inequality, we have

$$\begin{aligned}\|[\Delta]_T\|_\infty &\le \|(X_T^\top X_T)^{-1}X_T^\top\delta_j\|_\infty + \lambda\left\|\left(\frac{1}{n}X_T^\top X_T\right)^{-1}\text{sign}(\widehat{\theta}_T)\right\|_\infty\\ &\le \|(X_T^\top X_T)^{-1}X_T^\top\delta_j\|_\infty + \frac{\lambda}{\lambda_{\min}}.\end{aligned}$$

Since $\delta_j = X_j - X_T\Sigma_{TT}^{-1}\Sigma_{Tj}$, the first term can be expressed as

$$(X_T^\top X_T)^{-1}X_T^\top\delta_j = \widehat{\Sigma}_{TT}^{-1}\widehat{\Sigma}_{Tj} - \Sigma_{TT}^{-1}\Sigma_{Tj}.$$

Using the sub-multiplicativity of a matrix norm, we have

$$\begin{aligned}\left\|(X_T^\top X_T)^{-1}X_T^\top\delta_j\right\|_\infty &\le \left\|(\widehat{\Sigma}_{TT}^{-1} - \Sigma_{TT}^{-1})\Sigma_{Tj}\right\|_\infty + \left\|\widehat{\Sigma}_{TT}^{-1}(\widehat{\Sigma}_{Tj} - \Sigma_{Tj})\right\|_\infty\\ &\le \|\widehat{\Sigma}_{TT}^{-1} - \Sigma_{TT}^{-1}\|_2\left\|\Sigma_{Tj}\right\|_2 + \|\widehat{\Sigma}_{TT}^{-1}\|_2\cdot\sqrt{|T|}\left\|\widehat{\Sigma}_{Tj} - \Sigma_{Tj}\right\|_\infty.\end{aligned}$$

Applying Lemma 15 (Lemma 5 in Harris and Drton, 2013), if $\|\widehat{\Sigma}_{TT} - \Sigma_{TT}\|_\infty < \epsilon < \lambda_{\min}/|T|$, then

$$\|\widehat{\Sigma}_{TT}^{-1} - \Sigma_{TT}^{-1}\|_2 \le \frac{|T|\epsilon/\lambda_{\min}^2}{1 - |T|\epsilon/\lambda_{\min}}.$$

Hence, if $\|\widehat{\Sigma}_{TT} - \Sigma_{TT}\|_\infty < \epsilon < \lambda_{\min}/|T|$, we have

$$\|[\Delta]_T\|_\infty \le \left\|\Sigma_{Tj}\right\|_2 \frac{|T|\epsilon}{\lambda_{\min}(\lambda_{\min} - |T|\epsilon)} + \frac{1}{\lambda_{\min}} \cdot \sqrt{|T|}\epsilon + \frac{\lambda}{\lambda_{\min}}.$$

Applying the results of Lemma 18, we complete the proof. For a sub-Gaussian linear SEM,

$$\Pr\left(\|[\Delta]_T\|_\infty \le C_1|T|\epsilon + \frac{\lambda}{\lambda_{\min}}\right) \ge 1 - 4 \cdot \exp\left(\frac{-n\epsilon^2}{128(1+4s_{\max}^2)\max_j(\Sigma_{jj})^2}\right).$$

In addition for a $(4m)$-th bounded-moment linear SEM,

$$\Pr\left(\|[\Delta]_T\|_\infty \le C_1|T|\epsilon + \frac{\lambda}{\lambda_{\min}}\right) \ge 1 - 4 \cdot \frac{2^{2m}\max_j(\Sigma_{jj})^{2m} C_m(K_{\max}+1)}{n^m \epsilon^{2m}}.$$

where C_m is a constant depending only on m.

Applying the results of Lemma 18, we complete the proof. For a sub-Gaussian linear SEM,

$$\Pr\left(\|[\Delta]_T\|_\infty \le \epsilon'\right) \ge 1 - 4 \cdot \exp\left(\frac{-n}{128(1+4s_{\max}^2)\max_j(\Sigma_{jj})^2}\left(\frac{1}{C_1|T|}\left(\epsilon' - \frac{\lambda}{\lambda_{\min}}\right)\right)^2\right).$$

In addition for a $(4m)$-th bounded-moment linear SEM,

$$\Pr\left(\|[\Delta]_T\|_\infty \le \epsilon'\right) \ge 1 - 4 \cdot \frac{2^{2m}\max_j(\Sigma_{jj})^{2m} C_m(K_{\max}+1)}{n^m}\left(\frac{1}{C_1|T|}\left(\epsilon' - \frac{\lambda}{\lambda_{\min}}\right)\right)^{-2m},$$

where C_m is a constant depending only on m.

■

D.4 Proof for Lemma 13

Lemma 13 For any $r \in \{1, 2, ..., p-1\}$-th iteration, $j \in \{\pi_1, ..., \pi_{p+1-r}\}$, sufficiently small $\epsilon > 0$, and

- a sub-Gaussian linear SEM:

$$\begin{aligned}&\Pr\left(\left|\widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)}) - \mathrm{Var}(X_j \mid X_{S_j(r)})\right| < \epsilon\right)\\ &\ge 1 - 4 \cdot \exp\left(\frac{-n}{128(1+4s^2)\max_j(\Sigma_{jj})^2}\frac{\epsilon^2\lambda_{\min}^4}{(d+1)^2(\sigma_{\max}^2(\epsilon + 5\sigma_{\max}^2) + \epsilon\lambda_{\min})^2}\right).\end{aligned}$$

- a $(4m)$-th bounded-moment linear SEM:

$$\begin{aligned}&\Pr\left(\left|\widehat{\mathrm{Var}}(X_j \mid X_{S_j(r)}) - \mathrm{Var}(X_j \mid X_{S_j(r)})\right| < \epsilon\right)\\ &\ge 1 - 4 \cdot \frac{2^{2m}\max_j(\Sigma_{jj})^{2m} C_m(K_m+1)}{n^m}\left(\frac{\epsilon^2\lambda_{\min}^4}{(d+1)^2(\sigma_{\max}^2(\epsilon + 5\sigma_{\max}^2) + \epsilon\lambda_{\min})^2}\right)^{-m},\end{aligned}$$

where C_m is a constant depending only on m.

Proof

Consider a fixed $r \in \{1, 2, ..., p-1\}$-th iteration and $j \in \{\pi_1, ..., \pi_{p+1-r}\}$. For compactness in notation, subscripted j and r in parentheses are omitted, as done in Appendix A.

According to Proposition 14, the conditional variance of X_j given S satisfies the following inequality:

$$\mathrm{Var}(X_j \mid X_S) = \mathrm{Var}(X_j \mid X_T),$$

where T is the support of the solution θ^* in Equation (7).

Applying the above result and the Schur complement, we have

$$\begin{aligned}\left|\widehat{\mathrm{Var}}(X_j \mid X_S) - \mathrm{Var}(X_j \mid X_S)\right| &= \left|\widehat{\mathrm{Var}}(X_j \mid X_S) - \mathrm{Var}(X_j \mid X_T)\right| \\ &= \left|\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{jT}\widehat{\Sigma}_{TT}^{-1}\widehat{\Sigma}_{Tj} - (\Sigma_{jj} - \Sigma_{jT}\Sigma_{TT}^{-1}\Sigma_{Tj})\right| \\ &= \left|1/\widehat{\Omega}_{jj} - 1/\Omega_{jj}\right|,\end{aligned}$$

where Ω is the inversion of the covariance matrix for (X_j, X_T).

Then, it can easily be verified that

$$\left|\widehat{\mathrm{Var}}(X_j \mid X_S) - \mathrm{Var}(X_j \mid X_S)\right| < \epsilon \iff \left|1/\widehat{\Omega}_{jj} - 1/\Omega_{jj}\right| < \epsilon.$$

Applying Lemma 17 and Lemma 16, $|\widehat{\mathrm{Var}}(X_j \mid X_S) - \mathrm{Var}(X_j \mid X_S)| < \epsilon$, if

$$\|\widehat{\Omega} - \Omega\|_\infty \leq \frac{\epsilon}{\sigma_{\max}^2(\epsilon + 5\sigma_{\max}^2)}.$$

Now, applying Lemma 15 (Lemma 5 in Harris and Drton, 2013), it is a sufficient condition that $\widehat{\Sigma}$ satisfies the following:

$$\|\widehat{\Sigma} - \Sigma\|_\infty < \frac{\epsilon\lambda_{\min}^2}{(|T|+1)(\sigma_{\max}^2(\epsilon + 5\sigma_{\max}^2) + \epsilon\lambda_{\min})}.$$

Applying the results of Lemma 18, we complete the proof. For a sub-Gaussian linear SEM,

$$\begin{aligned}&\Pr\left(\left|\widehat{\mathrm{Var}}(X_j \mid X_S) - \mathrm{Var}(X_j \mid X_S)\right| < \epsilon\right) \\ &\quad\geq 1 - 4\cdot\exp\left(\frac{-n}{128(1+4s^2)\max_j(\Sigma_{jj})^2}\frac{\epsilon^2\lambda_{\min}^4}{(d+1)^2(\sigma_{\max}^2(\epsilon + 5\sigma_{\max}^2) + \epsilon\lambda_{\min})^2}\right).\end{aligned}$$

In addition for a $(4m)$-th bounded-moment linear SEM,

$$\begin{aligned}&\Pr\left(\left|\widehat{\mathrm{Var}}(X_j \mid X_S) - \mathrm{Var}(X_j \mid X_S)\right| < \epsilon\right) \\ &\quad\geq 1 - 4\cdot\frac{2^{2m}\max_j(\Sigma_{jj})^{2m}C_m(K_m+1)}{n^m}\left(\frac{\epsilon^2\lambda_{\min}^4}{(d+1)^2(\sigma_{\max}^2(\epsilon + 5\sigma_{\max}^2) + \epsilon\lambda_{\min})^2}\right)^{-m}.\end{aligned}$$

■

D.5 Proof for Lemma 15

Lemma 15 *(Lemma 5 in Harris and Drton, 2013) If $\Sigma \in \mathbb{R}^{q\times q}$ is a positive definite matrix, with minimal eigenvalue $\lambda_{\min} > 0$, and if $E \in \mathbb{R}^{q\times q}$ is a matrix of errors with $\|E\|_\infty < \epsilon < \lambda_{\min}/q$, then $\Sigma + E$ is invertible and*

$$\|(\Sigma+E)^{-1} - \Sigma^{-1}\|_\infty \leq \|(\Sigma+E)^{-1} - \Sigma^{-1}\|_2 \leq \frac{q\epsilon/\lambda_{\min}^2}{1 - q\epsilon/\lambda_{\min}}.$$

Lemma 15 is the same as Lemma 5 in Harris and Drton (2013), except for the bound $\|(\Sigma+E)^{-1} - \Sigma^{-1}\|_2$. Since the proof is directly from Harris and Drton (2013), we omit the proof.

D.6 Proof for Lemma 16

Lemma 16 *(Lemma 6 in Harris and Drton, 2013) If Σ is a covariance matrix for $X_{\pi_1,...,\pi_r}$ in a linear SEM (2), then the diagonal entries of Σ^{-1} satisfy $\Sigma_{jj}^{-1} \geq \frac{1}{\sigma_{\max}^2}$ where $\sigma_{\max}^2$ is the maximum error variance.*

Proof For any $j \in \{\pi_1, ..., \pi_r\}$ and $\mathrm{Pa}(j) \subset S := \{\pi_1, ..., \pi_r\} \setminus \{j\}$, recall that, for any variable X_j, its conditional variance is as follows:

$$\begin{aligned}\mathrm{Var}(X_j \mid X_S) &= \sigma_j^2 - \mathbb{E}(\mathrm{Var}(\mathbb{E}(X_j \mid X_S) \mid X_{\mathrm{Pa}(j)})) \\ &\leq \sigma_j^2 = \mathrm{Var}(\epsilon_j).\end{aligned}$$

This implies that

$$\Sigma_{jj}^{-1} = \frac{1}{\mathrm{Var}(X_j \mid X_S)} \geq \frac{1}{\mathrm{Var}(\epsilon_j)}.$$

Therefore, the diagonal entries of Σ^{-1} satisfy $\min(\Sigma_{jj}^{-1}) \geq 1/\max(\mathrm{Var}(\epsilon_j))$. ■

Lemma 16 guarantees that diagonal entries for the inversion of the correlation matrix are greater than, or equal to, $1/\max_j(\mathrm{Var}(\epsilon_j))$, which is a required condition for the following lemma.

D.7 Proof for Lemma 17

Lemma 17 *Let $A = (a_{jk})$ and $B = (b_{jk})$ be the 2×2 sub-matrices of an inverse covariance matrix and its estimated matrix. If A is positive definite with $a_{11}, a_{22} \geq a_{\min}$ and $\|A - B\|_\infty < \delta < a_{\min}/2$, then*

$$\left|\frac{a_{12}}{a_{11}a_{22}} - \frac{b_{12}}{b_{11}b_{22}}\right| < \frac{5\delta}{a_{\min}(a_{\min} - \delta)}.$$

Proof Without loss of generality, suppose $a_{12} \geq 0$. Since $\|A - B\|_\infty < \delta$,

$$\begin{aligned}\frac{b_{12}}{b_{11}b_{22}} - \frac{a_{12}}{a_{11}a_{22}} &< \frac{a_{12} + \delta}{(a_{11} - \delta)(a_{22} - \delta)} - \frac{a_{12}}{a_{11}a_{22}} \\ &= \frac{\delta}{(a_{11} - \delta)(a_{22} - \delta)} + a_{12}\left(\frac{1}{(a_{11} - \delta)(a_{22} - \delta)} - \frac{1}{a_{11}a_{22}}\right).\end{aligned}$$

Using $a_{11}, a_{22} \geq a_{\min}$ to bound the first term and $a_{12}^2 < a_{11}a_{22}$ to bound the second term, we obtain

$$\left|\frac{b_{12}}{b_{11}b_{22}} - \frac{a_{12}}{a_{11}a_{22}}\right| < \frac{\delta}{(a_{\min} - \delta)^2} + \sqrt{a_{11}a_{22}}\left(\frac{1}{(a_{11} - \delta)(a_{22} - \delta)} - \frac{1}{a_{11}a_{22}}\right).$$

Since the function $\sqrt{xy}(\frac{1}{(x-\delta)(y-\delta)} - \frac{1}{xy})$ is decreasing when $x, y \geq 0$, applying the condition $a_{11}, a_{22} \geq a_{\min}$, we have

$$\left|\frac{b_{12}}{b_{11}b_{22}} - \frac{a_{12}}{a_{11}a_{22}}\right| < \frac{\delta}{(a_{\min} - \delta)^2} + \left(\frac{a_{\min}}{(a_{\min} - \delta)^2} - \frac{1}{a_{\min}}\right) = \frac{1}{a_{\min}}\left(\frac{3\delta}{a_{\min} - \delta} + \frac{2\delta^2}{(a_{\min} - \delta)^2}\right).$$

Applying the condition, $\delta < a_{\min}/2$, we can see that

$$\left|\frac{b_{12}}{b_{11}b_{22}} - \frac{a_{12}}{a_{11}a_{22}}\right| < \frac{5\delta}{a_{\min}(a_{\min} - \delta)}.$$

■

Lemma 17 provides the error bound for a conditional variance given the error in inverse covariance matrix estimation. Its proof is analogous to the proof for Lemma 7 in Harris and Drton (2013) where a partial correlation bound is considered.

D.8 Proof for Lemma 18

Lemma 18 (Error Bound for the Sample Covariance Matrix) *Consider a random vector $(X_j)_{j=1}^p$ and suppose that its covariance matrix is Σ.*

- *Lemma 1 of Ravikumar et al. (2011): Suppose that $X_j^{(i)}$ are i.i.d. sub-Gaussian with proxy parameter $s_{\max}^2 \Sigma_{jj}$. Then,*

  $$\Pr\left(\left|\widehat{\Sigma}_{jk} - \Sigma_{jk}\right| \geq \zeta\right) \leq 4 \cdot exp\left(\frac{-n\zeta^2}{128(1 + 4s_{\max}^2)\max_j(\Sigma_{jj})^2}\right),$$

  for all $\zeta \in (0, \max_j \Sigma_{jj} 8(1 + 4s_{\max}^2))$.

- *Lemma 2 of Ravikumar et al. (2011): Suppose that $X_j^{(i)}$ are i.i.d. and there exists a positive integer m and scalar $K_m \in \mathbb{R}$ such that $\mathbb{E}(X_j^{4m}) \leq K_{\max}\Sigma_{jj}^{2m}$. Then,*

  $$\Pr\left(\left|\widehat{\Sigma}_{jk} - \Sigma_{jk}\right| \geq \zeta\right) \leq 4 \cdot \frac{2^{2m}\max_j(\Sigma_{jj})^{2m} C_m (K_{\max} + 1)}{n^m \zeta^{2m}},$$

  where C_m is a constant depending only on m.

Lemma 18 shows that the entries of the sample covariance satisfy an exponential-type tail bound with exponent $a = 2$, when samples are sub-Gaussian random vectors. Furthermore, Lemma 18 provides that the sample covariance entries satisfy a polynomial-type tail bound when samples are from random variables with bounded moments. Since it is the same as Lemmas 1 and 2 from Ravikumar et al. (2011), we omit the proof.

References

Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Peter Bühlmann et al. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013.

Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

Victor Chernozhukov, Wolfgang K Härdle, Chen Huang, and Weining Wang. Lasso-driven inference in time and space. *Available at SSRN 3188362*, 2019.

Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017.

Jianqing Fan, Shaojun Guo, and Ning Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems*, pages 6457–6466, 2017.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *The Journal of Machine Learning Research*, 14(1):3365–3383, 2013.

Jerry A Hausman. Specification and estimation of simultaneous equation models. *Handbook of econometrics*, 1:391–448, 1983.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Sven Klaassen, Jannis Kück, Martin Spindler, and Victor Chernozhukov. Uniform inference in high-dimensional gaussian graphical models. *arXiv preprint arXiv:1808.10532*, 2018.

Steffen L Lauritzen. *Graphical models*. Oxford University Press, 1996.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pages 745–752. ACM, 2009.

Radhakrishnan Nagarajan, Marco Scutari, and Sophie Lèbre. Bayesian networks in r. *Springer*, 122:125–127, 2013.

Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

Gunwoong Park and Yesool Kim. Learning high-dimensional gaussian linear structural equation models with heterogeneous error variances. *Computational Statistics & Data Analysis*, 154:107084, 2021.

Gunwoong Park and Youngwhan Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, 49(1):276–292, 2020.

Gunwoong Park and Hyewon Park. Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 158–166. PMLR, 16–18 Apr 2019a.

Gunwoong Park and Sion Park. High-dimensional poisson structural equation model learning via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 20(95):1–41, 2019b.

Gunwoong Park and Garvesh Raskutti. Learning large-scale poisson dag models based on overdispersion scoring. In *Advances in Neural Information Processing Systems*, pages 631–639, 2015.

Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *Journal of Machine Learning Research*, 18(224): 1–44, 2018.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15 (1):2009–2053, 2014.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Martin J Wainwright, John D Lafferty, and Pradeep K Ravikumar. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472, 2006.

Yi-Chi Wu, Ching-Sung Chen, and Yu-Jiun Chan. The outbreak of covid-19: An overview. *Journal of the Chinese Medical Association*, 83(3):217, 2020.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

Kun Zhang and Aapo Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–585. Springer, 2009a.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009b.